Prediction of Homing Pigeon Flight Paths using Gaussian Processes

Richard P. Mann

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2009

Richard P. Mann
Oriel College

DPhil Thesis
Michaelmas Term 2009

Prediction of Homing Pigeon Flight Paths using Gaussian Processes

---

Summary

Studies of avian navigation are making increasing use of miniature Global Positioning Satellite devices, to regularly record the position of birds in flight with high spatial and temporal resolution. I suggest a novel approach to analysing the data sets produced in these experiments, focussing on studies of the domesticated homing pigeon (*Columba Livia*) in the local, familiar area.

Using Gaussian processes and Bayesian inference as a mathematical foundation I develop and apply a statistical model to make quantitative predictions of homing pigeon flight paths. Using this model I show that pigeons, when released repeatedly from the same site, learn and follow a habitual route back to their home loft. The model reveals the rate of route learning and provides a quantitative estimate of the habitual route complete with associated spatio-temporal covariance. Furthermore I show that this habitual route is best described by a sequence of isolated waypoints rather than as a continuous path, and that these waypoints are preferentially found in certain terrain types, being especially rare within urban and forested environments. As a corollary I demonstrate an extension of the flight path model to simulate experiments where pigeons are released in pairs, and show that this can account for observed large scale patterns in such experiments based only on the individual birds' previous behaviour in solo flights, making a successful quantitative prediction of the critical value associated with a non-linear behavioural transition.

Publications

The work detailed in this thesis has been presented at the International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt) at the University of Mississippi, U.S.A. in July 2009 and at a Microsoft Research Cambridge seminar in June 2009. The MaxEnt presentation was accompanied by a paper published in the proceedings which is based on early work from Chapters 5 and 6 of this thesis.

- Mann, R., Freeman, R., Osborne, M., Garnett, R., Armstrong, C., Meade, J., Biro, D., Guilford, T. & Roberts, S. **Gaussian Processes for Prediction of Homing Pigeon Flight Trajectories.** *American Institute of Physics Conference Proceedings* **1193**, 360–367 (2009).

At the time of writing more developed work from these chapters is currently under consideration for publication in a peer-reviewed journal.

Contents

List of Figures

Acknowledgements

> "With the help of his friends and colleagues, Frodo passes through this ordeal, but discovers at the end that victory has no value left for him. While his friends return to settling down and finding jobs and starting families, Frodo remains in limbo; finally, along with Gandalf, Elrond and many others, he joins the brain drain across the Western ocean to the new land beyond." (*Lord of the Rings: an allegory of the PhD?*, David Pritchard)

Enormous thanks go to my supervisors, Tim Guilford and Steve Roberts. I gather it is rare to find a supervisor who is encouraging, full of ideas and genuinely interested in their students' work, so I feel especially lucky to have found two who were.
An honourable mention also to all the members of both OxNav and PARG. I would especially like to dedicate this thesis to two of my erstwhile colleagues:

- Mike, thanks for teaching me GPs, thanks for the millions of contentious shared G-Reader items, thanks for being an Aussie while England were winning the Ashes. The people of Crawley thank you for years worth of Morris dancing, milk drinking anecdotes. You're always welcome for Christmas at our house!

- Roman, thanks for pizza-as-bread, Glonass, aggregated diamond nanorods, whale turduckens and bacon salt. Thanks for alerting me to the doughnut in the foreground. Thanks for the Erdős number of 3. Thanks for giving me those documents the Russians wanted. Much obliged.

It is traditional at this point to thank your parents for their support. No one could ask for better than mine. Their help during the course of my doctorate was just one among a lifetime of countless acts of love and kindness.
This work was funded by the Engineering and Physical Sciences Research Council. Elements of the research contained in this thesis were also completed during internships at Microsoft Research Cambridge.

Chapter 1

Introduction

This thesis is a description of the development and application of a statistical model of homing pigeon flight paths and an argument in favour of using such a model to make quantitative predictions of flight path data to test biological hypotheses.

Flight paths are recorded using miniature Global Positioning Satellite (GPS) devices, which are attached to the pigeons' backs and record the position of the bird (in longitude and latitude co-ordinates) every second until the bird returns home, the device is retrieved and the data are downloaded. The most fundamental technical goal of this project is to determine a probability distribution for these data, such that a probability can be assigned to any observed flight path given information about the spatio-temporal restrictions on the birds' flight behaviour and the likely correlations between different flight paths. Doing so will allow Bayesian inference to be performed on the basis of observations made of real flight paths.

This work is necessarily cross-disciplinary. The majority of the original work in this thesis is mathematical in nature, describing the application of ideas from machine-learning to a novel zoological problem. In each chapter the results of the analysis are discussed in terms of the biological implications.

Relevant literature and ideas are reviewed from both a biological and a mathematical perspective. **Chapter 2** introduces the biological motivation for the work and reviews the relevant biological literature, leading with a brief overview of avian navigation theory and proceeding to a more detailed review of work associated with

visual navigation and navigation in the familiar area, with a particular focus on work since 1999 using GPS recording devices. **Chapter 3** and **Chapter 4** review the foundations of probability theory and the particular family of statistical models known as *Gaussian processes* (GPs). This constitutes the mathematical foundation of the thesis.

Chapter 5 describes how the flight path model is constructed. The framework of GPs is used to describe the probability of a flight path. A probability is also assigned to an unobserved *habitual route* around which multiple flight paths vary. This leads to a joint distribution of multiple flights that share a common habitual route, which ultimately allows the prediction of future flight paths from previously observed flight paths. Using the flight path model the development of predictability in initially naive birds released repeatedly from the same release site is observed and discussed with results from four different release sites.

Chapter 6 shows how the model can be used to identify waypoints used by a bird, by selecting for the most informative regions of the flight path and optimising the predictive power of the flight path model. Waypoints are identified from a number of birds and the discussion considers the spatial distribution of waypoints in relation to the underlying landscape.

Chapter 7 is concerned with collective navigation. With particular reference to a previous paper on experiments involving pigeons released as pairs, the flight path model is extended to describe the flights of pigeons released as pairs after extensive previous experience of the release site as individuals. The data from the previous experimental work is then used to assess the evidence in favour of a number of competing model adaptations. Samples are then drawn from the adapted model to simulate replication of the original experiments, which allows for replication of large scale phenomena observed in paired flight and accurate prediction of critical values associated with these phenomena.

1.1 Motivation

A decade after the first use of miniature GPS recording devices on navigating birds [Bürgi and Werffeli, 1999] the experimental use of tracking technology is reaching maturity. The field of avian navigation, especially experiments involving domestic pigeons, is approaching the point where tracking technology is reliable and cheap enough to be employed in the majority of studies.

The emergence of animal tracking as a source of high precision time-series data has attracted the attention of the machine-learning and statistics communities, and led to collaborations which have seen well established statistical techniques applied to analyse and filter these new data. Much of this work has consisted of established methodologies, such as noise reduction using the Kalman filter [Patterson et al., 2008] applied 'off-the-shelf', treating the animal tracking data as simply another time-series data source. Frequently these collaborations are characterised by a clear division of labour; zoologists acquire new data, submit it to statisticians for filtering and analysis and then interpret the analysis for its biological implications.

This thesis aims to blur this distinction, presenting a model designed from the ground up with particular biological data in mind. Using solid mathematical foundations this model will encode biological hypotheses directly, and will make quantitative predictions of real movement data based on those hypotheses. By doing so it will perform a greater task than describing the data or filtering it for easier interpretation; it will allow the observed data to adjudicate in favour of alternative hypotheses in a quantitative manner, to allow us to judge which best explains and, more importantly, which best predicts the results of experiments.

Biology is a science of complex phenomena. When the underlying processes that produce observable data are sufficiently complex it can be easy, and tempting, to explain how a favoured theory could have produced the observed effect, to explain or describe the observations in terms of the hypothesis. While it may be comforting to conclude that observations support a favoured hypothesis, or to find a hypothesis

that appears consistent with the observations, to do so is an abdication of the duty to test hypotheses. That a given theory *can* explain the data shows only that the particular hypothesis *could* be true. Alternative theories could be proposed that also fit the observations. The better test is whether the theory predicts the results of new experiments, and crucially whether it does so better than the alternatives. Therefore it is important, even when modelling complex phenomena, to make concrete predictions that can be tested against the predictions of other theories.

Chapter 2

Avian Navigation and Project Aims

2.1 Introduction

Every year large areas of the world play witness to an extraordinary feat of navigation as huge numbers of migratory birds arrive from distant parts of the earth, escaping harsh climatic conditions or taking advantage of favourable feeding conditions. In extreme cases these migrations may take the bird half way round the world on each trip. The Manx Shearwater for instance, migrates annually between the British Isles and the coast of Argentina, a round trip of some 40,000 kilometres. This migration is a staggering feat simply in terms of endurance for such a small creature under its own power and we are only just beginning to understand how it is performed [Guilford et al., 2009]. Greater than the endurance challenge however is the difficulty of accurately navigating between relatively small areas separated by such enormous distances. Remarkably these birds often return to precisely the same burrow used in previous years when breeding.

Humans have long recognised and utilised the unique power of avian navigation. The combination of a natural navigational system, combined with the large distance travel permitted by flight, allows birds to migrate to exploit seasonal variations in food and weather, and allows humans to train birds to carry messages to distant locations, a method that was in widespread use for many centuries before the advent of modern

telecommunications. Recreational use of these abilities continues in pigeon racing. Media interest in new discoveries relating to avian navigation reveal a wide interest in the phenomenon. This interest has developed from an observation of an ability, through practical use of domesticated birds, to the modern science of navigational studies. Scientists now systematically ask *how* birds manage these feats rather than what use they can be put to:

> Such a conspicuous phenomenon as the long-distance flights of birds has profoundly penetrated into man's consciousness, and it is a very simple further step to ask how they find their way [Kramer, 1952]

This chapter will introduce the current state of knowledge in the field, with particular reference to the canonical experimental species — the domesticated pigeon, (*Columba livia*). The pigeon has been the basis for the majority of experimentation in this field due to its expedient properties. As recognised in Wallraff's detailed review of the field [Wallraff, 2005], pigeons have acted as a 'laboratory rat'. Originally domesticated for food and trained for years to send messages or for racing there is a large body of knowledge on the rearing and keeping of pigeons. They can be kept in relatively large numbers in simple lofts. They are social animals and display an urge to return to the home loft throughout the year.

This thesis occupies a niche position within the broad field of avian navigation, focusing on a completely new methodology for analysing high density, high precision data that has only been available within the last decade. As such, understanding of this work does not require an in depth knowledge of the wide scope of avian navigation research, nor does it depend on fully exposing the most current controversies that currently define each specific area. It is, however, important to set both this work, and the work upon which it directly depends, within the broader context of avian navigation research. Therefore this review will present an overview of the principle themes of research which define the field, which will not only set context but also elucidate the meaning of various terminology that will be used throughout the rest

of the thesis. Areas of current disagreement within the literature will be noted where appropriate, but it is not the intention to make a judgement on these controversies where they do not impact work specific to this thesis. This will be followed by a more detailed review of the specific areas that are critical to the understanding of the work in later chapters.

Excluded from this chapter is a review of how pigeons and other birds navigate collectively. The research in this field is more specifically relevant to Chapter 7 only, therefore the review is presented in that chapter in Section 7.2.

The review presented here necessarily owes a large debt to previous extensive reviews by Papi [1990], Schmidt-Koenig [1990], Wiltschko and Wiltschko [1996], Holland [2003], Wallraff [2004] and especially Wiltschko and Wiltschko [2003] and Wallraff [2005], which provide a fuller historical and thematic review of the field of avian navigation, in its entirety, than can be presented here.

2.2 Definitions

Navigation is a broad term that may encompass a wide range of animal behaviours and strategies that enable the animal to successfully move from a starting location to a desired goal. This thesis will be examining a subset of avian navigation known as *homing*, where the final destination is always the bird's home. To achieve the goal of reaching home a bird may employ a variety of strategies. In familiar locations the bird may have sufficient direct experience of the environment to navigate using recognisable *landmarks*, which may be visual or other sensory cues that are *geostationary* — fixed in the landscape. This type of navigation has been commonly referred to as *pilotage* [Wallraff, 2005], since in its simplest form the bird *pilots* from one landmark to the next in a chain to reach home. When landmarks are used in this fashion they constitute *waypoints* — fixed locations that the bird physically visits on the journey home. See Section 2.7 for more details about navigation in the familiar area.

In unfamiliar locations, areas which the bird has little or no direct experience of, navigation by known landmarks is not possible. Instead, to reach home the bird must have an alternative means to determine the direction of home and to maintain a consistent flight in that direction. The term *navigation* is often used to mean the ability to determine the home direction without direct experience of the immediate environment and without cues emanating from home [Baker, 1984], but here I follow the terminology used by Wallraff [2005] and refer to this as *true navigation*, using *navigation* more broadly to represent all goal oriented movement. In research on true navigation a *Map* and *Compass* model [Kramer, 1953] has formed a dominant paradigm that provides the intellectual context for most experiments. This model supposes that true navigation approximates the manner in which a human uses a map and a compass. A *map* is a means of identifying position relative to the goal. A *compass* is a means for orienting to, and maintaining, the correct heading to reach that goal. By modelling true navigation in this manner, research has been able to focus on revealing the sensory mechanisms used to replicate the functionality of the map and the compass, and this has proved a successful paradigm. Nonetheless, as Wallraff [2005] cautions:

> While speaking about compasses and maps, however, we should be aware that we are using anthropomorphic metaphors.

Therefore we should be careful not to impose too strictly the limitations of actual human maps and compasses upon the birds' internal representations of the world.

Whilst the natural familiar area will vary from bird to bird as a result of its own foraging patterns, specific areas can be induced to become part of the familiar area through experimental intervention — by displacing birds to *release sites*, where they are released to fly home. Repeated displacement to the same release site can incorporate that site into the familiar area (see next section).

2.3 Experimental procedure and analysis

Evidence for and against theories of avian navigation has traditionally been acquired through displacement experiments, where birds are artificially transported to a pre-selected release site. In some cases this site will be entirely unknown to the bird while in others it may be a previously visited site or within the bird's familiar area. This displacement is typically accompanied by an experimental manipulation of some sensory input. The effect of the manipulation is then assessed via a set of standard metrics. The proportion of successful homing and the average flight duration indicate how efficiently the bird is able to home. So called 'vanishing bearings' are used to assess whether there is a directional bias in the flight. This is done by following the bird using binoculars for as long as possible and noting the bearing at which it was last sighted — hence the terminology. These metrics, especially the actual practice of taking vanishing bearings, have now been to some extent overtaken by the use of GPS tracking.

The phenomenon of 'release site bias' [Keeton, 1973] poses a serious obstacle to analysing data from release experiments. Site specific properties can cause artifacts that appear to indicate either a positive or negative result from an intervention. Wallraff [1996] (theses 4 and 5) discusses the dangers of making inference from a single release site. An example in this thesis would be the release site at Weston Wood, north-east of Oxford (see Figure 5.1). The release site is close to a major road that runs in the direction of the loft. Previous work has shown that even naive birds released from unfamiliar sites will be attracted to prominent visual features of the landscape – such as villages – even without recognising them from previous experience [Kiepenheuer, 1993, Wallraff, 1994]. Many birds fly along this major road even on their first release from the site. Naively interpreted, it could appear that these birds have an excellent estimate for the homeward direction when in fact they could be simply attending to very local cues. This illustrates why results, especially striking results that contradict previous experiments, obtained from releases at a single site

should be treated with some caution or scepticism.

2.4 Compass Mechanisms

The compass mechanism is understood to be primarily driven by two sensory systems. The first, and dominant mechanism is the *Sun Compass* [Kramer, 1953], using the position of the sun in conjunction with an internal clock to determine the current bearing. The second mechanism is a magnetic sense, the existence of which is generally accepted although the actual sensory system is the subject of dispute.

2.4.1 Sun compass

The sun, being enormously distant and extremely prominent, offers a directional cue that, at a fixed point in time, changes very slowly with position on the earth. As such it is an ideal cue on which to base a compass mechanism. However, because the position of the sun in the sky changes through the day it can only be used to determine a bearing in conjunction with an independent measure of time. The independent measure of time dictates the expected position of the sun in the sky. Bearings can then be obtained relative to this fixed direction.

The sun's position in the sky changes in two angular dimensions over the course of the day. The sun rises and sets — changing in elevation, or the *zenithal* angle. In non-equatorial areas of the globe, or outside of the equinox, the tilt of the earth's axis relative to its orbit also cause the sun's *azimuthal* angle (the angle in the horizontal plane) to change strongly throughout the day, progressing from East to West. In equatorial conditions at equinox the sun also progresses from east to West, but does so directly overhead, changing only the zenithal angle.

It is the azimuthal angle that birds use to orient to a bearing. There is no evidence that birds attend to the elevation of the sun [Schmidt-Koenig, 1990]. The azimuthal angle progresses predictably throughout the day at a fixed position on the earth's surface. This progression is highly dependent on latitude and season, as the

length of day varies. At the equator during equinox the sun passes almost directly overhead. Thus the azimuthal angle passes directly from East to West without any continuous variation of the azimuthal angle. There is evidence that birds continue to use the sun compass in equatorial regions at equinox [Schmidt-Koenig, 1990], but that a separate mechanism is used (potentially the magnetic compass, see below) when the sun is close to directly overhead. Wiltschko and Wiltschko [1981] published results suggesting that the bird must learn the relationship between the sun's azimuthal angle and the time of day, thus suggesting that the variation over latitudes and seasons can be adjusted for by recalibrating that relationship.

To use this sun compass requires independent knowledge of time, for which the birds must have an 'internal clock'. Evidence for the use of the sun compass comes from altering the apparent angle of the sun, or by manipulating this internal clock through 'clock-shifting'. Clock-shifting involves confining the bird to a closed loft without external light, where the hours of artificial light are controlled to simulate different daylight hours. Over time the bird's internal clock adjusts to this artificial day, perturbing the relationship between internal time representation and the position of the sun. The same effect is experienced as 'jet-lag' in humans when travelling across widely spaced longitudes.

Migratory birds, caged during their natural migration season, show restlessness which is directed towards the required migratory heading [Kramer, 1950]. By using an arrangement of mirrors to alter the apparent angle of the sunlight entering the cage, this restlessness can be predictably directed towards a new angle [Kramer, 1952]. This demonstrates the impact of the sun on orientation. Clock-shift experiments on homing pigeons, which naturally orient towards the home loft without intervention, have shown that this impact is seen during navigation as well as in confinement. Numerous studies have shown that clock-shifting produces predictable deflections in the bird's initial bearing after release (see Schmidt-Koenig [1990]). This supports the Map and Compass model, indicating that the initial bearing is determined through an independent compass mechanism and not from pattern recognition of the location

(e.g. the relative positions of a number of recognised visual features might provide directional cues). It is also a clear validation of the sun compass — the predictability of the deflection after clock-shifting shows that the mechanism operates as supposed. Experiments have demonstrated the use of the sun compass over the full range of homing, from migratory birds held in cages showing directed restlessness within the cage [Kramer, 1952], through long distance pigeon homing experiments (see review and references by Schmidt-Koenig [1990]), down to pigeons released less than one kilometre from their home loft [Wilkinson et al., 2008, Armstrong et al.].

2.4.2 Magnetic compass

The full predictable deflection in the initial bearings of clock-shifted birds released from unfamiliar sites suggests that no secondary compass mechanism is available that would partially compensate for the perturbation. Despite this pigeons are found to home successfully from unfamiliar sites even in overcast conditions [Keeton, 1971]. In addition, clock-shifted birds near the equator are found to orient successfully when the sun is directly overhead and cannot provide directional information [Schmidt-Koenig, 1990]. This suggests the existence of a secondary compass mechanism that is strictly secondary, i.e. one that is only referred to in situations when the sun is unavailable for obtaining a bearing. It is now generally accepted that a magnetic sense allows the geo-magnetic field to be used as secondary compass system, but that this is strictly secondary to the sun compass, such that the magnetic compass does not compensate for errors in the sun compass but is only used when the sun is not visible.

Orientation with reference to the earth's magnetic field is a trait common to many animals [Wiltschko and Wiltschko, 2005], and is of course the basis for the compass used by humans. It is unsurprising, therefore, that birds use this mechanism as well. Aside from isolated, anomalous regions near large ferromagnetic deposits the earth's magnetic field is a very consistent source of accurate directional information.

The magnetic sensory system can potentially be disturbed by attaching ferromagnets to birds, which if sufficiently strong may overwhelm the ability to sense

the relatively weak geo-magnetic field. The review by Keeton [1971] reports that pigeons released with such magnets attached became disoriented when the sky was overcast and the sun was not visible. Pigeons released, with magnets attached, during sunny conditions generally showed little or no disorientation (see also Wallraff [2005]), confirming the dominant role of the sun compass. Wiltschko and Wiltschko [1996] also suggest that use of the magnetic compass may be responsible for the eventual return to the loft of clock-shifted pigeons, which, it is claimed, generally return home within one day, before the sun compass could be recalibrated. Without details of the flight path from tracking experiments it is difficult to judge whether this theory is correct, or whether an alternative strategy search as a random search for familiar territory is responsible for this return.

The use of the magnetic compass is particularly prominent in migratory birds. Unlike a human compass, the magnetic compass in migratory birds is controlled by the inclination of the local magnetic field relative to the gravitational force vector, what Wiltschko and Wiltschko [1996] call the 'inclination compass' (see Wiltschko and Wiltschko [1996], figures 1, 3 and 4 and Wallraff [2005], figure 6.6). Because the inclination of the magnetic field is inverted in the Southern Hemisphere relative to the Northern Hemisphere, while the direction of the gravitational vector remains directly downwards, the same rule can encode for poleward or equatorward travel in both hemispheres. Wiltschko and Wiltschko [1996] argue, based on the results of Wiltschko and Gwinner [1974] and Beck and Wiltschko [1982], that this allows migratory patterns to be 'hard-coded' as instinctive, rather than learnt, responses, since most migrations are polewards or equatorwards. The instinct to move towards the equator when conditions become too harsh may be a very early evolutionary development in birds that has since proved adaptive in both hemispheres.

The biological mechanism underlying the magnetic compass is highly contested. Some argue in favour of a sensory system based on a magnetite element in the bird's beak, that is ferromagnetic and thus subject to a physical force in the presence of magnetic fields, while others argue for the magnetically modulated activation of cer-

tain photoreceptor molecules in the retina, known as *cryptochromes*. See Zapka et al. [2009] for a recent summary and contribution to that debate.

2.5 Map Mechanisms

In the Map and Compass model the bird senses its location in space, thus determining the correct bearing home before consulting the compass to fly along that bearing. Determining its position in space, especially in unfamiliar locations is a more complex process than determining direction. This is reflected in the relatively poor understanding we have regarding the map mechanisms compared to the compass mechanisms.

2.5.1 Olfaction

The olfactory map (reviewed in detail by Papi [1990], Wallraff [2004]) is the most strongly evidentially supported hypothesised mechanism for determining the bird's position in space in unfamiliar areas. Birds that are made anosmic through olfactory nerve section, anaesthetisation or occlusion of the nostrils [Papi et al., 1971, 1972, Hartwick et al., 1977, Papi et al., 1980] or birds that are denied access to directional air flows from outside the loft [Baldaccini et al., 1974, Ioalè et al., 1978] are significantly less efficient at homing from previously unvisited release sites. Moreover, if the direction of air flows are manipulated before entering the loft the birds are observed to be predictably deflected at the release site, leaving in a direction consistent with the angle of deflection [Baldaccini et al., 1975, Ioalè et al., 1978, Waldvogel et al., 1978, Kiepenheuer, 1978]. The birds' initial bearings after release can also be predictably perturbed by exposing them to odours from a different site [Benvenuti and Wallraff, 1985]. Ioalè et al. [1990] elegantly combined the directional breezes at the loft with a false odour experiment, in which pigeons were exposed to benzaldehyde from one direction at the loft; pigeons subsequently exposed to benzaldehyde at release were initially oriented in the opposite direction to the artificial breeze at the loft. Recent

work by Jorge et al. [2009] challenges the clarity of these results. Extending false-release experiments such that birds are taken to one site, then exposed to artificial odours unrelated to that site, before being released at a second site after being made anosmic appears to show that birds react as if released at the first site, despite never receiving true odours at that site. This, it is suggested, shows that the presence of the odours has an activational effect that 'switches on' a non-olfactory positional mechanism. This appears inconsistent with the predictable deflections achieved by rotating the airflow at the home loft and it remains to be seen how these results can be reconciled.

This is strong evidence that olfaction is a consistently used map mechanism, at least at previously unseen release sites where visual recognition is unavailable. Anosmic birds home more successfully from previously used release sites [Benvenuti et al., 1973, Hartwick et al., 1977] and even from novel release sites within the familiar area [Wallraff and Neumann, 1989], suggesting that familiar locations can be recognised without olfactory cues. The evidence of clock-shift (see Section 2.4.1) experiments from familiar sites suggests decreasing but still-present deflection with increasing familiarity [Wallraff, 1996], suggesting that the olfactory map and sun compass mechanisms are partially compensated by an orientation mechanism based on site recognition, potentially through the spatial pattern of visual landmarks (see Section 2.5.2). Experiments with clock-shifted anosmic birds released at familiar sites suggest that the recognition of a familiar site can provide the bird not only with a spatial location (the 'map') but also with directional cues (the 'compass') [Bingman and Ioalè, 1989]. This is supported by the observed flight paths of clock-shifted birds released at familiar sites [Biro et al., 2007], which show that the effect of clock-shift is altered, from a directional bias to a displacement relative to the non-clock-shifted flights, which remains stable throughout the flight.

The required directionality of airflow at the loft suggests that the birds associate odours or trace atmospheric compounds with directions of airflow and therefore fly in the appropriate direction to the loft when displaced to a location where those odours

or compounds are abundant. Where the concentration of a compound varies sufficiently smoothly in space the bird may also be able to extrapolate from its experience in the familiar area. For example, if a given odour becomes consistently stronger the further west the bird travels during foraging it may associate strong concentrations of that odour with the need to fly east. By using compounds that vary along orthogonal directions a bird could potentially pinpoint its location in space by using the concentration of each compound as a spatial axis. To learn this gradient map a bird would need to be able to move around its home to sample the concentration of each compound and assess its variation in space. Support for this mechanism comes from experiments on birds reared in windless loft conditions. Birds that are reared in windless lofts and are not allowed out of the loft to sample the air for three months fail to successfully home even if they are subsequently fully exposed to directional air flows before being released, suggesting that the olfactory map associated with directional air flows is formed in early development [Gagliardo et al., 2001]. However, of those birds that were deprived of directional air flows when young, those which are subsequently allowed to fly freely around the loft subsequently home more successively than those simply exposed to directional air flows later in development [Ioalè et al., 2008]. This indicates that after three months the primary olfactory map can no longer be developed, but a secondary map based on olfactory gradients can be.

The directional scatter of the initial bearings of birds released at unfamiliar sites is much larger than at familiar sites [Wallraff, 1996], suggesting that if olfactory information is used to deduce position it has a low resolution. This is relevant to navigation in the familiar area, where birds are observed to attend to features with very high resolution, indicating that olfaction alone can not account for this precision.

2.5.2 Vision

The use of vision to determine position intuitively seems very likely, although this may be an anthropomorphism since vision is the primary human sensory system. However, for a time vision was perceived to be largely irrelevant and potentially even

disruptive to successful homing. Schmidt-Koenig and Schlichte [1972] found that pigeons wearing frosted contact lenses were able to home successfully from up to 130 kilometres from home to within a few hundred metres of the loft, despite being unable to recognise previously learnt artificial landmarks at a distance of 6 metres. This paper clearly showed that while *some* pigeons were able to home successfully to the vicinity of the loft with impaired vision, the experimental disruption of vision caused major disruption to the birds' decisions at release (many refused to fly or flew a short distance before stopping) and their homing within the vicinity of the loft (most birds did not enter the loft once within the vicinity). Moreover there were dramatic changes in flight behaviour, as the authors note:

> The usual behaviour of the birds upon release and at the loft was also drastically altered in experimental birds. Upon release, many experimental birds refused to fly, hovered, or crash-landed nearby; others hit wires, trees, or other obstacles. All birds that did fly did so in a peculiar way...Experimental birds usually arrived at the loft rather high in the sky, cautiously hovering down, a few hitting, most others missing, the loft. [Schmidt-Koenig and Schlichte, 1972]

Schmidt-Koenig and Walcott [1978] indeed acknowledge in a later paper that navigation near the loft is probably driven by a mechanism that the frosted lenses interfere with — a visual system. However, despite these caveats, the growing body of evidence of the primacy of the olfactory map, both from the ability of pigeons to return to the vicinity of home with impaired vision, along with a continued impact of clock-shifting in the vicinity of the loft [Graue, 1963, Keeton, 1974, Schmidt-Koenig, 1979, Füller et al., 1983] – which indicated that birds still made reference to the compass even in sight of very familiar areas – effectively sidelined research into visually mediated navigation.

Work by Braithwaite and Guilford [1991] marked the return of vision as an experimentally supported homing mechanism. Pigeons, previously familiarised with a

release site through six training releases, were placed in specially constructed boxes for five minutes prior to a subsequent experimental release. In one group these boxes were transparent, affording the bird a view of the landscape. In the other group the boxes were opaque, denying the bird any detailed visual input from the landscape prior to release. When kept in the transparent boxes, the birds subsequently reached the home loft significantly faster than when kept in the opaque boxes, in individual comparisons. This finding was confirmed by further releases at a loft in Germany [Braithwaite, 1993] (the original experiment was conducted in England), demonstrating robustness to site specific effects. Further work showed that pigeons released at unfamiliar sites did not home faster when allowed to preview the landscape [Braithwaite and Newman, 1994], indicating that the effect was due to recognition of the site, as opposed to a potential psychological effect of being denied vision or because the birds were able to 'plan' the first section of their flight while visually previewing. Burt et al. [1997] replicated the same experimental procedure but with a redesigned pre-release box to ensure full access to olfactory cues, showing that disruption to olfaction is not necessary in order for birds to use visual cues. Biro et al. [2002] eventually demonstrated that the improved navigation times were due to reduced route tortuosity in the early part of the flight by using miniature GPS devices attached to the birds (see Section 2.6 below). This was a further demonstration that behaviour was changed in the vicinity of the recognised release site, rather than over the entire flight duration.

Further evidence for the role of vision in improving navigation comes from experiments on clock-shifted birds (see Section 2.4.1). Birds that are clock-shifted to induce deflections in the initial bearings after release, through manipulation of the sun compass, are significantly less deflected when allowed to preview a familiar release site [Wallraff et al., 1999]. Evidence from GPS tracked birds demonstrates clearly that birds previously familiarised with a specific release site, which are observed to repeatedly follow the same route home from that site (see Section 2.7.1), will continue to follow that route even when clock-shifted [Biro et al., 2007]. These findings are disputed by Wiltschko et al. [2005], who claim that the initial bearings of clock-shifted

birds are not affected by site familiarity. They do, however, acknowledge the effect of site familiarity upon homing times. Wiltschko et al. [2005] review several prior experiments which support the familiarity effect on initial bearings, finding fault in the experimental procedure of each. It is hard though to reconcile their objections with the quantity of evidence from GPS tracking that now supports both visual navigation and reduced deflection in clock-shifted birds from familiar sites.

Vision was also implicated in potentially influencing the initial bearings of birds released from unfamiliar sites, where recognition of the site would be impossible. Wallraff [1994] demonstrated that birds released from a wide spread of unfamiliar locations showed a correlation between their vanishing bearings and the locations of nearby villages and forests — birds were more likely to fly over built up areas (villages) and showed a preference for open areas rather than wooded areas. This showed that visual identification of the landscape structure could lead to a consistent bias in initial bearing without a relation to the bird's homeward direction. This is also a reminder that a navigating bird will likely use many criteria in its search for a route home, balancing not only efficiency but also certainty of return and risk of predation.

2.5.3 Magnetic map

The idea of a map mechanism based on the geomagnetic field has a long history, dating back to work by Yeagley [1947, 1951] who showed that position could be determined through the intersection of true latitude, as measured by the Coriolis force, and the geomagnetic latitude, as determined by the strength of the local magnetic field. While theoretically possible this would require both the magnetic field and the Coriolis force to be measured with great precision. There is no strong evidence at present to suggest birds can sense their spatial location using the geomagnetic field, and Wallraff [2005] is witheringly critical of the theoretical speculation regarding magnetic maps in the absence of direct evidence of their existence (see Wallraff [2005], page 81). Some suggestive evidence is found in the behaviour of birds released within geomagnetic

anomalies, such as Dennis et al. [2007], but these studies tend to show that anomalies cause unexpected responses, which could equally well be explained by stress induced on the bird by the disturbed state of the magnetic compass. Dennis et al. [2007], for example, shows that pigeons released in an anomaly tend to align their flights either parallel or perpendicular to the local magnetic field. While surprising, this does not constitute a mechanism for localisation. Unexpected responses to unusual stimuli such as the magnetic anomaly could potentially be a symptom of sensory 'stress'. If the stimulation to a bird's sensory inputs is sufficiently outside of normal bounds this could trigger a response intended to escape the sensation, or alternatively trigger more general 'fear' responses. See Wallraff [2005], page 79, for more discussion of the effect of stressful stimuli.

A prerequisite for a convincing demonstration of the existence of a magnetic map would be a predictable directional or positional deflection as the result of a magnetic disturbance. So far this has not been demonstrated.

2.6 Global Positioning Satellite technology

Traditional measures of navigational behaviour are centred on information available at release and collection of the bird, such as the initial bearing and the time taken to reach home. These leave the majority of the flight unprobed, since vanishing bearings typically correspond to the birds orientation after a maximum of 2 kilometres of flight, whilst experimental flights can extend for over 100 kilometres. This left an enormous potential source of information about birds' navigational strategies unexplored.

Efforts have been made to investigate the entirety of the flight using progressively more advanced technology throughout the twentieth century [Schmidt-Koenig and Walcott, 1978, Bramanti et al., 1988, Dall'Antonia et al., 1993, Bonadonna et al., 1997]. The use of direction recorders [Bonadonna et al., 1997], which provided the most accurate tracking technology before the use of GPS, allowed an accuracy within approximately five to ten degrees of the true heading over the whole flight (estimated

Figure 2.1: A pigeon from the Oxford Field Station, wearing a GPS position logger on its back.

from Bonadonna et al. [1997], equating to a maximum error of roughly 50 metres per kilometre flown. While these provided valuable insights into in-flight behaviour at the time of use they have since been eclipsed by the extra-ordinary precision offered by the use of Global Positioning Satellite (GPS) technology, which uses orbiting satellites to determine position on the earth's surface to within a resolution of approximately five metres at all times. (see Freeman [2009] for details).

The introduction of GPS devices small enough to be carried by a bird in flight (see Figure 2.1 for an example) has revolutionised the experimental study of avian navigation. GPS data has confirmed many previous findings, notably demonstrating the power of the sun-compass even in the familiar area [Biro et al., 2007]. It has enabled the detailed study of simultaneous path data from multiple birds to look for co-operative behaviour [Biro et al., 2006b, Freeman, 2009]. The biggest impact of using GPS technology has been in the study of navigation within the familiar area, where it has revealed the dominant role of vision through the observation of 'road-following' behaviour and the discovery of habitual routes (see Section 2.7)

The nature of data collected from tracking devices has prompted the development of a range of new analysis tools. Algorithmically defined metrics have been created to pick out certain features of the data, so as to detect particular forms of behaviour. A good example of this is the metric of *first-passage time* [Fauchald and Tveraa, 2003]. This measures the amount of time a tracked animal remains in a given area, from the first moment it enters the area until the first moment it leaves. It has been suggested that this provides an algorithmic way to separate the recorded path into different behavioural types, such as foraging or travelling [Barraquand and Benhamou, 2008]. As the density and accuracy of tracking data has increased it has become more necessary to employ algorithmic data analysis to make full use of all the information available in the data.

The desire to extract the most useful information from tracking data has prompted a collaboration between the fields of zoology and machine-learning. Faced with the question of how to best analyse high density time-series data, elements of the animal

navigation community have turned towards machine-learning as a field built on creating algorithms to deal with this exact type of data. Some of these collaborations can be simply an exchange of algorithms. For example, the Kalman filter [Kalman, 1960], a well established tool of the engineering and machine-learning sciences for decades, can be used to reduce the error from a low precision tracker just as well as it can be used to reduce the noise in an electrical signal or to improve the accuracy of a rocket's navigation systems (see Patterson et al. [2008] for examples). Equally, generic 'change-point' algorithms designed to detect discontinuous changes in signals can be used to interpret first-passage time measurements from foraging animals to identify the boundaries between different behavioural types [Barraquand and Benhamou, 2008].

In some cases there has been a closer collaboration, integrating ideas from machine-learning more deeply in the analysis. Rather than simply employing previously developed tools 'off the shelf', machine-learning has been used to describe the data in an entirely new way. In the avian navigation literature, an ambitious attempt to use machine-learning to probe tracking data more deeply can be seen in the work of Roberts et al. [2004] and Guilford et al. [2004]. These studies developed and applied a new metric of *positional entropy*. This was derived by considering the local predictability of the track, in essence calculating how much new information each recorded position provided. The higher the amount of information from the new fix, the lower the predictability. If the new position was very predictable – if the bird was moving in a straight line, for example – the new position provided very little information, since the result of the measurement could already be predicted from earlier positions. This metric, therefore, was both designed to analyse tracking data but also rooted very much in the conceptual framework of machine-learning, built on ideas of information theory and probability. A *Hidden Markov model* [Rabiner, 1990], another established machine-learning method, was then employed to segregate the paths into a number of behavioural states, thus defining distinct behaviours solely in terms of an information-theoretic view of the recorded paths. This metric was subsequently used

by Lau et al. [2006], to demonstrate the coupling between the visual landscape and behavioural states (see Section 2.7), and Freeman [2009], to demonstrate the coupling between the behavioural states of birds when flying together (see Section 7.2).

The increasing use of GPS technology in tracking animals means that algorithms designed specifically to analyse this particular form of data will correspondingly increase in importance. While generic algorithms can be applied to good effect in the right circumstances, optimal use of the data will increasingly require models created specifically with particular zoological problems in mind. This then is the principal motivation of this thesis, to build a model from first principles, specifically designed for the experimental system being analysed (see Section 2.9).

Where are the limitations of tracking technology? Wallraff [2005] points out that contemporary GPS tracking technology is expensive, and devices must be collected to download the recorded data. This makes them unsuitable for experiments where birds may not return to the loft, since in such cases expense will be incurred and no data collected. As with all technology though the cost of the devices has continually reduced since their development and more groups are able to afford to use tracking technology in a wider range of experiments every year. As of 2009 a suitable commercially available device, suitable for use with pigeons, costs less than £50 and GPS tracking experiments are becoming the norm in avian navigation groups worldwide. The increasing ubiquity of GPS technology in cameras, cars, computers and mobile phones is continuing to push down prices. As Wallraff [2005] also concedes, satellite telemetry (whereby data is downloaded from the device 'on-the-fly', potentially through the mobile phone network) could be a solution to the necessity of collecting the birds.

In addition to the trend towards smaller, cheaper GPS devices, which will open up the use of tracking technology to more groups and on smaller animals, there are more ambitious projects to extend the range and precision of tracking technology further. The ICARUS (Improving Cooperation for Animal Research Using Space) initiative published a white paper [Wikelski and Rienks, 2008] in which it argued for

using either a new or existing satellite system at a lower orbit than the GPS satellites. This would increase signal strength at the ground a require less power for ground-based devices. These devices could then become smaller since they would require less battery capacity. As the white paper indicates, this would be an enormous benefit, since most animals are small (the distribution of animal sizes is skewed), and animal weight determines the size of device that can be carried without deleterious effects. Thus a huge number of new animals would be available for tracking experiments.

With the eventual completion of the European Union's Galileo project (a European version of the American GPS system) improved accuracy can be expected, both because the new system is designed for greater accuracy and because, with more satellites in orbit, fewer position fixes will need to be taken from a sub-optimal number of satellites. It is also expected that fee-paying users will be able to obtain the same accuracy as military users, which is typically significantly better than the standard accuracy available for free.

2.7 Navigation in the Familiar Area

In addition to the Map and Compass model, which describes how birds can navigate home from unfamiliar areas, it has been hypothesised that birds also have a separate mechanism, which Baker [1982] calls a 'familiar area map', that allows them to home from familiar locations [Holland, 2003]. The set of all familiar locations constitutes the familiar area.

The familiar area is defined as being an area of which the bird has direct physical experience. This can be from independent foraging flights or due to navigating from release sites chosen by an experimenter. Direct experience of an area implies that the bird can recognise its position as being one previously visited, rather than inferring its location from one of the map mechanisms described above.

The familiar area, defined as above, will naturally vary between individual birds. There are few examples of birds being tracked during natural foraging behaviour,

though the Oxford group has documented two cases of birds, whilst wearing trackers, spontaneously flying approximately 25 kilometres from the home loft before returning (Tim Guilford, personal correspondence). Experiments by Biro et al. [2006a] suggest that pigeons are able to recognise familiar landmarks at sites 25 kilometres from the home loft and use them to navigate home. Therefore 25 kilometres represents a convenient lower bound on the potential extent of the home range.

For releases within this range there is now strong evidence in favour of *pilotage* — navigation through visual recognition of geo-stationary landmarks, which the bird 'pilots' between, visiting each in turn. As well as the evidence given above in favour of a non-specific visual element in navigation the deployment in recent years (since 1999) of GPS recording devices has provided a strong support for the pilotage hypothesis. The most persuasive evidence comes from the observation of 'route fidelity' [Biro et al., 2002, Biro, 2002, Biro et al., 2004, Meade et al., 2005, 2006, Biro et al., 2006a, Armstrong et al., 2008] (and see also Bonadonna et al. [1997] for pre-GPS evidence of route fidelity). Birds that are released repeatedly from the same release site within 25km of the home loft are, in the vast majority of cases, observed to form and follow 'habitual routes', typically varying by less than 250 metres along that route. These routes are also characterised by regions of extremely low variation ($\sim$ 10-20 metres) which are highly suggestive of memorised geo-stationary locations — *waypoints*. An example of this phenomenon is shown in Figure 2.2, which shows five successive recorded flight paths from a single individual, after extensive training (15 previous releases) from the release site. The distance between the paths is consistently on the order of 100 metres. Of the potential map mechanisms described above only vision is likely to have sufficient resolution for the bird to remain so close to previously memorised locations. This is supported by the findings of Schmidt-Koenig and Schlichte [1972]. Although birds were able to navigate from unfamiliar locations to within a few hundred metres of the loft without visual cues they were not able to enter the loft itself, suggesting that the non-visual cues they relied on have a resolution too low to explain the extremely high precision return to particular locations

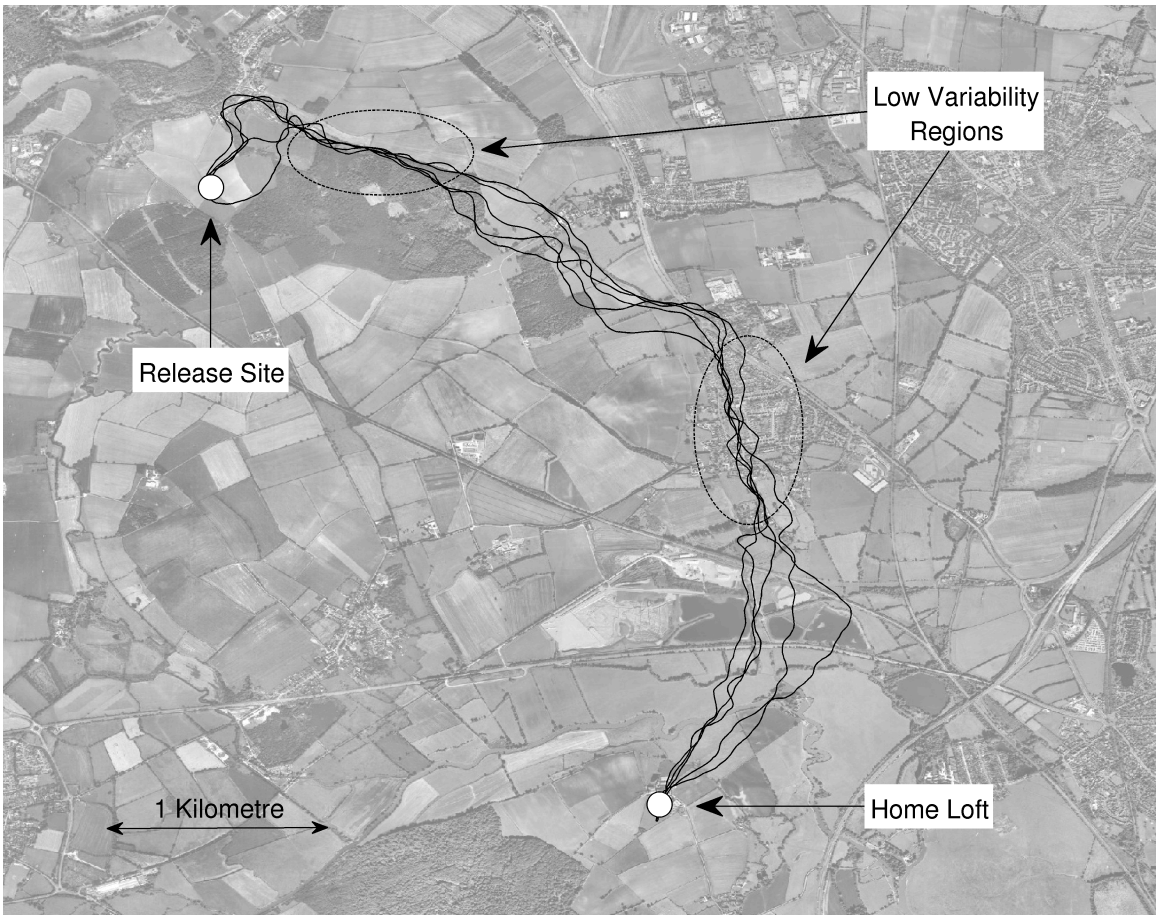


Figure 2.2: An example of habitual route following. Five successive flights from one individual released at a familiar site. Circled by dashed lines are regions of particularly low variability, suggestive of a salient feature in the landscape.

observed in birds trained at familiar release sites. Strong evidence for a visual system underpinning the birds' loyalty to their routes is found in the concurrence of many of the routes with prominent visual features in the landscape, particularly roads [Lipp et al., 2004, Biro et al., 2004].

2.7.1 Habitual routes

Over a series of releases from an initially novel release site birds have been observed to form habitual routes from a variety of different release sites, from a wide spread of compass bearings and at up to 25 kilometres distance around the field station loft at Oxford University (England) [Biro et al., 2002, Biro, 2002, Biro et al., 2004, Meade

et al., 2005, 2006, Biro et al., 2006a, Armstrong et al., 2008]. These routes tend to be highly idiosyncratic, though the extent of idiosyncrasy differs between sites. The individuality in this choice suggests attention to a personal set of landscape features, potentially point-like landmarks, that allow the bird to recapitulate the same route repeatedly.

Although it is expected that visual cues must underlie the memorisation of the habitual route, it is not yet clear what the particular features are that birds attend to in flight. Lipp et al. [2004] and Biro et al. [2004] observed a tendency for pigeons to recapitulate routes over prominent linear features in the landscape, particularly major roads, with some birds even switching roads at major junctions. Since strong linear features in the landscape almost never line up perfectly with the home direction, there must be an instinct for following these cues that includes a tolerance for moving in the wrong direction. It is possible that, in addition to their potential use as navigational cues, such features are favourable to remain close to for other reasons such as reduced predation risk. They may also share visual characteristics with pre-urban elements of the landscape, such as the boundaries of natural terrain types, that would have dominated the visual landscape before the emergence of urbanisation.

Beyond this observation it has been impossible to determine the entire class of visual features which landmarks are composed of. The primary reason for this difficulty has been the inability to manipulate the landscape on the scale of these experiments. One cannot construct a featureless landscape on which to impose a small set of features, so as to determine whether the birds attend to them. Instead, landmarks must be identified from observational studies, by releasing birds in the natural environment and determining, *post-hoc*, the salient areas of the landscape.

Detecting salient regions of the landscape from observed flight paths presents an ideal opportunity to apply machine-learning methods. Classifying the data into salient regions represents a classic task in signal processing. In principle the task is straight-forward — given a known set of landmarks, and a series of flight paths that are known to attend to these landmarks, the task is to identify something unique

about the pattern of the flight paths in the vicinity of the landmarks. By looking for the same pattern in other flight paths more landmarks can be identified. Unfortunately, no such data set exists. In the absence of so called 'labelled data' – data with known features, in this case landmarks – this type of learning is impossible. One of the primary aims of this thesis will be to create an algorithmic method for identifying landmarks, without the need to look for specific localised patterns in the data, such that the landscape can be classified into regions of interest for further analysis in determining which visual cues are most important to the birds. This is discussed further in Chapter 6.

Lau et al. [2006] showed what an objective analysis of what constitutes landmarks might look like. Based on the observation of linear-feature following they used a standard edge-detection algorithm [Canny, 1986] to decompose an image of the landscape around the release sites and the home loft into a binary 'edge image'. The edge image showed where there were strong contrasts in the original image, picking out features such as roads, hedgerows, forest boundaries, etc. They then showed that the observed flight paths from a number of release sites flew over a greater number of these 'edges' than would be expected, suggesting that these were an important element of the birds' navigational memories. They subsequently used the positional entropy metric of Roberts et al. [2004] to show that edges were associated with changes in behavioural states, again demonstrating an entirely objective link between algorithmically defined aspects of the landscape and of the flight path. To truly understand how the landscape affects navigation, rather than simply demonstrating that it does, one would need to construct a predictive model that incorporated elements of the landscape as a factor. Ongoing work by Yukie Ozawa and Graham Taylor of Oxford University has focused on simulating flight paths based on attraction to a particular landscape feature (hedgerows) and shows some early promise at making accurate predictions of real flight paths, although this has not yet been generalised to more than a single release site experiment [personal correspondence with Yukie Ozawa]. Currently this approach suffers from the need to hand-label the landscape character-

istic of interest and an overly mechanistic and deterministic approach to simulating the flight path. Combining the hypothesised landscape influence with a more flexible stochastic approach to prediction could lead to a clearer picture of how specific visual features influence flight paths.

The formation of habitual routes is queried by Wiltschko et al. [2007]. Their experiments, releasing birds from a site separated from the loft by a principally urban environment (Frankfurt, Germany) did not show the characteristic sign of habitual route formation — reducing distance between successive flight paths. This was a surprising result in the light of the number of examples of habitual route formation produced by the Oxford group. One explanation may be that the birds used by Wiltschko et al. [2007] were not initially naive — they had prior experience of the release site before the experiment in question. This may mean that the learning phase was already complete and therefore no more improvement would be observed. The published GPS tracks do not, however, look like idiosyncratic habitual routes in the form seen repeatedly around Oxford. Another potential explanation is that the particular, highly urban, landscape affected the birds ability to learn a route. Although Wiltschko et al. [2007] proposed that the urban environment presented an ideal chance for route learning, due to the high visual information content of the landscape, it may be that pigeons are not able to process or memorise visual information of such detail. Over broad scales urban landscapes may all look very similar, preventing recognition of a particular site. Armstrong et al. [2008] performed experiments including a release that forced birds to cross urban Oxford to reach the home loft. In that study the birds were found to reduce their between-flight distance somewhat less quickly, and to a slightly greater asymptotic distance than a comparison site in a more rural environment, although this result was not very strong and was potentially confounded by the existence of a large road at the more rural site heading in the direction of the home loft.

Since the existence of habitual routes is a fundamental component of the subsequent work in this thesis, it is worth establishing how robust this phenomenon is.

Once formed, loyalty to the habitual route is consistent under a variety of experimental perturbations.

Displacement

Biro et al. [2004] demonstrated that birds released from novel locations displaced up to 1.5 kilometres perpendicular to the previous established habitual route would return to the habitual route downstream (meeting the original route closer to the loft than at the point closest to the displaced site). This was later shown to apply for displacements of up to 3 kilometres [Biro et al., 2006a]. Habitual route recapitulation from displaced release sites was also observed by Armstrong et al. [2008].

Compass perturbation

Adjustment of the sun compass through clock-shifting in birds with previously established habitual routes results in the bird following the habitual route but displaced a small distance towards the direction of the compass disturbance [Biro et al., 2007], demonstrating both the robustness of the habitual route but also that the bird has not entirely neglected the input of the compass. This deflection around a previously established habitual route is observed even at releases within 900 metres of the home loft [Wilkinson et al., 2008, Armstrong et al.], suggesting that the birds are unable to neglect their directional sensory inputs, even in extremely familiar areas.

Pigeons are not prevented from forming habitual routes under disruption of the magnetic sensory system, through the attachment of powerful magnets (relative to the geo-magnetic field) to their heads during training [Meade et al., 2005]. Therefore it is unlikely that the memorisation route can depend on, or be disturbed by changes to, the magnetic compass.

Flights with other birds

To examine conflicting navigational aims Biro et al. [2006b] released birds in pairs after they had established habitual routes from the same release site as individuals. The

majority of individuals returned to their individual habitual route when subsequently released individually. See Chapter 7 for further details.

2.8 Conclusion

Exactly how pigeons and other birds navigate from both familiar and unfamiliar locations, either to home or to a distant migratory target, remains an unsolved problem. Overwhelming evidence exists for the use of the sun, and the geo-magnetic field, as directional cues ('compasses'), and for the use of olfactory information and visual cues to determine position ('maps'). Pigeons display a remarkably robust ability to find home under sensory disruption that suggests a large redundancy in these mechanisms — removal of one system is rarely sufficient to prevent the bird returning to its home loft.

The evidence of very structured and repeated flight paths in the familiar area suggests a hierarchical system of information, where the bird primarily uses familiar visual cues when they are available, but relies on a Map and Compass system dominated by the olfactory map and the sun compass (with the magnetic compass in a secondary role) when they are not.

Navigation in the familiar area is mediated, at least in part, through pilotage – successively visiting fixed waypoints – as shown by the low variability in repeated flight paths by experienced birds – a phenomenon that is robust to extensive experimental perturbation. The exact nature of these waypoints and how they are recognised is not fully understood, and the difficulty in objectively identifying them remains an obstacle to this understanding.

2.9 Aims

The study of avian navigation has entered an exciting new era of experimentation through the use of new technology that allows high precision tracking of birds in flight. This provides the opportunity for new kinds of experiment and analyses but

also presents the challenge of vast new data sets. Recorded flight paths represent high density data series composed of extremely correlated positional data. One approach to using this data is to extract metrics, low-dimensional properties of the data. Examples would include 'virtual vanishing bearings' (determining the flight bearing a pre-determined distance from the release site), efficiency (the total distance travelled as a multiple of the straight path between release site and the loft) and fidelity (the distance between paths). A potential alternative approach is to begin viewing the flight paths themselves as the fundamental unit of observation. With GPS equipment becoming ever more widely used, the basic measurement from release experiments will soon be 'the flight path'. To make inference from measurements one must be able to make statistical comparisons between the measurements made and those expected under a given hypothesis. The aim of this thesis is to demonstrate a method for making such comparisons, for linking models of flight behaviour to the probability of observed flight paths.

This thesis demonstrates the development and application of a statistical model for the purpose of studying avian navigation. The model creates a framework for learning a distribution of flight paths, correctly assigning probabilities to observed paths and making predictions about the future flight paths of navigating birds based on previously seen data. This forms the basis of a new form of analysis for this type of zoological data — model comparison. By framing hypotheses as extensions to this framework we can assess the descriptive and predictive power of alternate hypotheses and objectively decide on their relative merits by using Bayesian model comparison criteria.

The aim of the work is to determine the distribution of high-dimensional GPS data, so as to predict the outcomes of future experiments (where the outcome is the resulting flight path). At present such data is largely confined to experiments within the familiar area, since GPS devices have historically been too costly to deploy in experiments where birds are not guaranteed to return, and the bird must be collected to retrieve the data. Hence the work in this thesis is focused on experimental releases

within a roughly 10 kilometre radius of the home loft. All data was collected by the Oxford group and is previously published [Meade et al., 2005, Biro et al., 2006b, Armstrong et al., 2008].

Since the focus is primarily on navigation in the familiar area, the outstanding problem this thesis will address is the identification of waypoints. Prior to this I will construct a model that constitutes an adaptable distribution of flight paths. This will be constructed in such a way that it can 'learn' from previous observations to more accurately describe subsequent flight paths. This will be tested by observing the habitual route learning process of initially naive pigeons from the new perspective of *increasing predictability* (Chapter 5). It will then be demonstrated how adjusting the model to use only the most informative subset of the previous observations can make superior predictions, and conversely how optimising the predictive power of the model can allow selection of the most informative data. This, it will be argued, constitutes an objective method for identifying waypoints or landmarks.

As a corollary to this principle goal, a further chapter will also investigate applications of the same model to group navigation, demonstrating how previously observed characteristics of paired-release experiments can be understood in terms of the distribution of flight paths from each individual bird.

Chapter 3

Probability Theory

3.1 Introduction

In order to make robust model-based inferences about such flight paths we need to make forecasts, assess in a principled manner how accurate the forecasts are and perform model selection in a consistent manner so as to fairly compare alternative hypotheses that explain our observations. Probability theory and probabilistic inference offer a principled and robust framework for such requirements.

This chapter introduces the rules and methods of probability theory that will be applied in later chapters. It begins by defining terminology and explaining the philosophical interpretation of probabilities as 'degrees of belief' in propositions. The use of probability theory to make inferences based on observed data, to select between competing models or hypotheses and to determine optimal model complexity follows. Finally a full treatment of Bayesian linear regression is presented, both to provide a case study for the ideas explored in this chapter and to lead towards the description of Gaussian processes in the next.

This chapter attempts to provide a thorough treatment of probability theory without excessive technical details, as is appropriate for this type of cross-disciplinary work; the reader should be able to take from this chapter everything required to understand the use of probability theory in later chapters.

3.2 Definitions

This thesis uses the term *probability* in the Bayesian sense, to mean a 'degree of belief' in a system of extended logic (see Section 3.3 for alternative definitions). This means that a probability represents the extent to which a proposition can be believed to be true, based on the information available. In classical Aristotelian logic a proposition may be either *True* or *False*. In the extended logic of probability theory we have beliefs about whether or not propositions are true, ranging from 0 (definitely false) to 1 (definitely true). This chapter will generally follow the approach and notation of Jaynes from his definitive account of probability theory [Jaynes, 2003], which in turn follows the structure of the work by Cox [Cox, 1946, 1961], which forms the basis for the 'degree of belief' interpretation. All probabilities will be explicitly conditional, since the correct degree of belief in any proposition depends on the information available. The catch-all conditioning term, I, will represent information not otherwise explicitly used, such as the structure of the problem or previous experience of the likely range of values for some variable.

The probability of a proposition, A, is denoted as $P(\,A \mid I\,)$. The negation of a proposition is represented as $\neg A$. Some propositions refer to the value of *random variables*. In the Bayesian interpretation a random variable can be any numerical quantity about which we are uncertain. A typical proposition takes the form $X = x$, which denotes represents the statement 'the variable X takes the value x'. In general the variable will be denoted by the uppercase and the value it takes by the lowercase. The probability, $P(\,X = x \mid I\,)$, will often be shortened, simply as $P(\,x \mid I\,)$, when it is clear which variable the probability refers to.

Probability density functions are defined on propositions over continuous variables and are denoted by the lower case. The *probability density* is defined such that the probability that the value of a variable lies within some interval is the integral of the probability density over that interval,

$$\int_a^b p(\,X = x \mid I\,)\,\mathrm{d}x = P(\,a < X < b \mid I\,)\,. \tag{1}$$

This and the later chapters will make extensive use of the Gaussian probability density function, which is defined in Appendix A.

3.3 Foundations

Equivalent rules for the manipulation of probabilities can be derived from differing philosophical viewpoints. In the frequentist interpretation probabilities represent the proportion of times a given event will occur over a large number of repeated trials. For example, the probability of a coin toss landing on heads is the proportion of times a heads will occur when a coin is tossed a large number of times. Formally, probabilities represent the proportion of times the event occurs in the limit as the number of trials approaches infinity.

In the Bayesian interpretation, probabilities are taken to indicate the *degree of belief* of some agent, be that a person or a machine, in a given proposition. For example, if $C = H$ represents the proposition that the next toss of a coin will be a Heads, then $P(\, C = H \mid I \,)$ represents the probability, or degree of belief, that that proposition is true and that a heads will occur. Note that the probability is conditioned on an external factor, I. Here I represents our information or knowledge about the world. For example, I might contain the information 'most coins are fair' or 'I do not know which way the coin might be biased', and therefore $P(\, C = H \mid I \,)$ would be one-half.

Cox [1946] proposed that a well-behaved, quantitative measure of belief must obey the same rules as traditional frequency probability — namely the product rule and the sum rule. However, having reached this point through consideration of the plausibility of propositions, a Bayesian is free to apply them without recourse to hypothetical frequencies of results in large numbers of hypothetical experiments. This will allow us to address questions that cannot be posed as the result of a repeatable experiment, such as 'is there life on Mars?', or more pertinently, 'which hypothesis is most likely?'.

The use of probability theory to reason about degrees of belief can be justified in a variety of alternative ways. See Halpern [2003] for a review of such justifications.

3.3.1 Rules for manipulating probabilities

The Product Rule

The product rule specifies how to calculate the joint probability of two propositions given the conditional probability of one given the other.

$$P(\,A, B \mid I\,) = P(\,A \mid B, I\,)\,P(\,B \mid I\,) = P(\,B \mid A, I\,)\,P(\,A \mid I\,) \tag{2}$$

The Sum Rule

The sum rule states that either A or $\neg A$ must be true, and that both cannot be true simultaneously. Thus the probabilities of both must add to unity.

$$P(\,A \mid I\,) + P(\,\neg A \mid I\,) = 1 \tag{3}$$

Marginalisation

Using the sum and product rules we can derive the important result of *marginalisation*. From the product rule:

$$P(\,A, B \mid I\,) + P(\,A, \neg B \mid I\,) = [P(\,B \mid A, I\,) + P(\,\neg B \mid A, I\,)]\,P(\,A \mid I\,)\,. \tag{4}$$

And noting from the sum rule that the elements in square brackets must equal one we have:

$$P(\,A \mid I\,) = P(\,A, B \mid I\,) + P(\,A, \neg B \mid I\,)\,, \tag{5}$$

which can be generalised to the case where B may take many values

$$P(\,A \mid I\,) = \sum_b P(\,A, B = b \mid I\,) = \sum_b P(\,A \mid B = b, I\,)\,P(\,B = b \mid I\,)\,. \tag{6}$$

This allows us to remove unknown or unwanted variables from the analysis, while fully accounting for the uncertainty that this introduces.

Bayes' rule

Rearrangement of the product rule leads to to the famous theorem of Bayes

$$P(\,A \mid B, I\,) = \frac{P(\,B \mid A, I\,)\,P(\,A \mid I\,)}{P(\,B \mid I\,)} \tag{7}$$

This rule allows for the reversal of a conditional probability, in this case calculating how strongly the truth of B implies the truth of A based on the degree to which the truth of A would imply the truth of B. Particularly we should note that this probability depends on how likely A is *a priori* before we learn about B, through the quantity $P(\,A \mid I\,)$. We call this the *prior* probability of A and denote the updated probability of A, in the knowledge of B, the *posterior* probability of A.

Bayes' rule allows us to *update* our belief in some proposition in the light of changing information, and thus forms the basis for performing inference within the Bayesian framework.

3.4 Bayesian Inference and Model Selection

Bayesian inference is concerned with obtaining the probability of some proposition given the information available. Typically we will be interested in propositions of the form: 'the value of X is x', where X may be something we are trying to predict (e.g. the value of the stock market tomorrow) or a parameter in a model (e.g. the CO_2 absorption of the ocean in a climate model). In Bayesian terms, data, parameters and predictions are all interchangeable — they are all quantities about which we are more or less uncertain and about which we aim to make inferences.

Denote the probability density that the random variable, X, takes particular value, $X = x$, by $p(\,X = x \mid I\,)$, which will be frequently shortened to $p(\,x \mid I\,)$ when it is clear which variable we are discussing. Henceforth I will in general use probability densities, since most problems involve data and parameters that take continuous values, though the same results apply for discrete probability distributions.

Suppose we are in possession of some data, $D = d$, which is informative about

X. Then Bayes' rule tells us the *posterior* distribution of X,

$$p(\, X = x \mid D = d, I\,) = \frac{p(\, D = d \mid X = x, I\,)\, p(\, X = x \mid I\,)}{p(\, D = d \mid I\,)}. \tag{8}$$

The first quantity in the numerator on the right hand side is the *likelihood*. This is usually specified by our model — we state the probability of observing a particular value of the data as a function of the possible values of X. The second quantity is the *prior* distribution of X. This specifies our beliefs about the value of X before we observe D.

The denominator on the right hand side is a normalising constant and can be calculated by considering the numerator for all values of X since,

$$p(\, D = d \mid I\,) = \int_x p(\, D = d \mid X = x, I\,)\, p(\, X = x \mid I\,)\, \mathrm{d}x. \tag{9}$$

3.4.1 Incorporating uncertainty

Consider a model, M, that aims to predict the value of some future data, D_*. This model contains a parameter, Θ, the value of which will determine the probability of seeing $D_* = d_*$. We can learn about Θ by incorporating other data, $D = d$, we have previously observed, but except in very special cases the data we have observed will still leave us uncertain about the value of Θ. From Bayes' rule the posterior distribution of Θ is given by

$$p(\, \Theta = \theta \mid D = d, I\,) = \frac{p(\, D = d \mid \Theta = \theta, I\,)\, p(\, \Theta = \theta \mid I\,)}{p(\, D = d \mid I\,)}. \tag{10}$$

A naive approach to the problem of predicting D_* would be to find the optimum value of Θ, according to some heuristic, and use that to determine the probability distribution over D_*. In classical statistics this would take the form of maximising $p(\, D = d \mid \Theta = \theta, I\,)$ and is known as Maximum Likelihood Estimation (MLE). While this is an apparently sensible heuristic, this is to *confuse the inverse*. That the data is *most likely* given a particular value of the parameters does not imply that those parameter values are the most probable, since not all values of the parameters are

equally probable *a priori*. Bayes' rule shows that the probability distribution of the parameter values is a product of the prior probability as well as the likelihood. MLE is often proposed as an 'objective' procedure, since there is no need to specify any prior belief on the distribution of the parameters. This is, however, an illusion. MLE implicitly argues that all parameter values are equally likely *a priori*, which is clearly false in cases where those parameters have been estimated previously, or where physical constraints impose limits on the values the parameters may take. In the event of a Maximum Likelihood Estimate disagreeing substantially with previously acquired knowledge about the parameter value, one might be inclined to believe the result to be faulty and disregard or repeat the experiment. This is a *subjective* and potentially inconsistent use of prior information that would be applied with greater clarity by the appropriate use of a prior distribution.

Perhaps then one should find the optimum value of Θ by maximising the posterior probability. This avoids some of the more egregious consequences of the maximum likelihood method and is also widely used. However, it is easy to see that this may also lead to some unfortunate consequences. The maximum value of the posterior may be somewhat meaningless if the distribution is wide or skewed. There is no guarantee that the peak of the distribution must lie close to the bulk of the probability. Indeed, the peak of a probability density function does not represent any real optimum; the density is only meaningful as a probability when integrated over an interval. Therefore an isolated peak does not represent the most probable value. Moreover, the actual peak value is not invariant under arbitrary changes of co-ordinate system; the choice of parameter value should not depend on which co-ordinates we choose to describe the problem in. Also, by choosing an optimum value we overestimate our certainty from that point forward if we fool ourselves by believing we have determined the 'true' value of Θ. While there are undoubtedly some cases where considerations of computational resources or mathematical complexity demand the selection of a single optimum value for a given parameter this should not be the default action.

The mathematically consistent approach is to avoid *optimising* at all. Equation

(6) already provides the answer to our problem. We wish to know the probability of new data, conditioned on our model and the data we have already observed, $p(\, D_* = d_* \mid D = d, M, I\,)$, without having to specify a value for the unknown parameter. Therefore we marginalise over the possible values of the parameter.

$$\begin{aligned} p(\, D_* = d_* \mid D = d, M, I\,) = \int_\theta & p(\, D_* = d_* \mid D = d, M, \Theta = \theta, I\,) \\ & \times p(\, \Theta = \theta \mid D = d, M, I\,)\, \mathrm{d}\theta \end{aligned} \tag{11}$$

Which, by application of Bayes' rule leads to

$$\begin{aligned} p(\, D_* = d_* \mid D = d, M, I\,) = \int_\theta & p(\, D_* = d_* \mid D = d, M, \Theta = \theta, I\,) \\ & \times \frac{p(\, D = d \mid \Theta = \theta, M, I\,)\, p(\, \Theta = \theta \mid M, I\,)}{p(\, D = d \mid M, I\,)} \mathrm{d}\theta. \end{aligned} \tag{12}$$

By marginalising we correctly incorporate our uncertainty in Θ into our uncertainty in D_*.

3.4.2 Model selection

Suppose we wish to use a data set, D, to determine which of two hypotheses, H_0, H_1 is correct. The objective way to make such a decision is to assess the relative probabilities of each hypothesis in the light of the data. Bayes' rule tells us that we should assess both the *marginal likelihood*, of each hypothesis – the probability of the data in the light of each hypothesis being true – and the prior probabilities.

$$\frac{P(\, H_1 \mid D, I\,)}{P(\, H_0 \mid D, I\,)} = \frac{p(\, D \mid H_1, I\,)}{p(\, D \mid H_0, I\,)} \frac{P(\, H_1 \mid I\,)}{P(\, H_0 \mid I\,)} \tag{13}$$

When framed as mathematical models each hypothesis may contain elements of uncertainty, represented by adjustable parameters, Θ. We incorporate this uncertainty by marginalising over the unknown parameters by integration,

$$p(\, D \mid H, I\,) = \int p(\, D \mid H, \Theta, I\,)\, p(\, \Theta \mid I\,)\, \mathrm{d}\Theta \tag{14}$$

This marginalisation allows us to accurately assess the evidence in favour of different hypotheses. Rather than comparing the best possible predictions of each hypothesis

(by choosing the best set of parameters), we compare them over the full range of predictions they make, based on how well we can estimate what those parameters are. This is particularly useful in choosing models of the optimum complexity, where over-fitting is a common problem.

Jeffreys [1939] gave a personal interpretation of the Bayes factor, later endorsed by Penny et al. [2004]. This relates the numerical value of the Bayes factor to the strength of the evidence. Jeffreys considered any value of the Bayes factor below three to be insubstantial evidence and definitely not decisive in favour of the more probably hypothesis. However, a Bayes factor is in fact no more or less than what it mathematically represents — the relative probabilities or 'odds' of two proposed hypotheses. Therefore a Bayes factor of three simply means that one hypothesis is three-times more probable than the other.

Model Complexity

A particular example of model selection is in the choice of an appropriate level of model complexity. Often we have a choice of models of varying complexity to fit the same data. A simple example of this would be in curve fitting — should we look for a linear fit or does the data support a higher order polynomial? In classical statistics this choice often comes down to the judgement of the statistician. For MLE methods a *penalised likelihood* may be used, introducing a penalty term dependent on the number of adjustable parameters to penalise more complex models. This is justified by reference to *Occam's razor*, the oft-quoted principle that when two hypotheses explain the data equally well, the simpler hypothesis should be favoured.

Parsimonious models tend to make better predictions since they capture the most important aspects of the data without trying to fit unexplainable variation. This is often utilised by ignoring a portion of the data at the model fitting stage and judging model comparison by attempting to predict this excluded data, a process referred to as *cross-validation*. The model that most accurately predicts the excluded data is judged to be the best.

Bayesian methods allow us to avoid many of the difficulties associated with model order selection. Bayesian analysis provides a mathematically rigorous version of Occam's razor through the marginalisation of model parameters. To see this consider the difference between Bayesian marginalisation and maximum likelihood (or maximum *a posteriori*). The space of free parameters in a model can be seen as representing the range of possible 'worlds' that model can describe. In maximising the likelihood we are in effect free to choose the one instance of the model that best fits our data. The more free parameters available, the more possible instances we have to choose from and the more likely we are to find one that matches the real observed data. The extreme case of this would be a model specifying exactly the value of each datum as a delta function. In this case we would have as many free parameters as we have data and would be able to pick a 'perfect' fit that explains all the observations. However, the predictive power of such a model would typically be zero, since there is no relationship between the values of the data, so the next observed datum can take any value. By contrast, in Bayesian analysis we marginalise over the parameters. In effect this means we have to average how well each of the possible instances of the model specified by the parameters fits the data. A model with more parameters than necessary will have a larger set of possible data sets that can be explained and the probability mass will accordingly be spread thinly. Therefore models that fit the data well will have a low weighting. Too simple a model will describe too few possible data sets to find an appropriate fit to the observed data. But in between there will lie a model of the correct complexity that fits the data well but parsimoniously. Figure 3.1 indicates graphically the intermediate ground between over-simplicity and over-complexity.

Example: polynomial curve fitting

Polynomial curve fitting provides a canonical example of selecting the correct model complexity. Consider a data set, $[\boldsymbol{x}, \boldsymbol{t}]$. We wish to model $\boldsymbol{x}$ as being a polynomial function of $\boldsymbol{t}$ with additive zero-mean Gaussian noise of variance, ϵ^2. However, we are

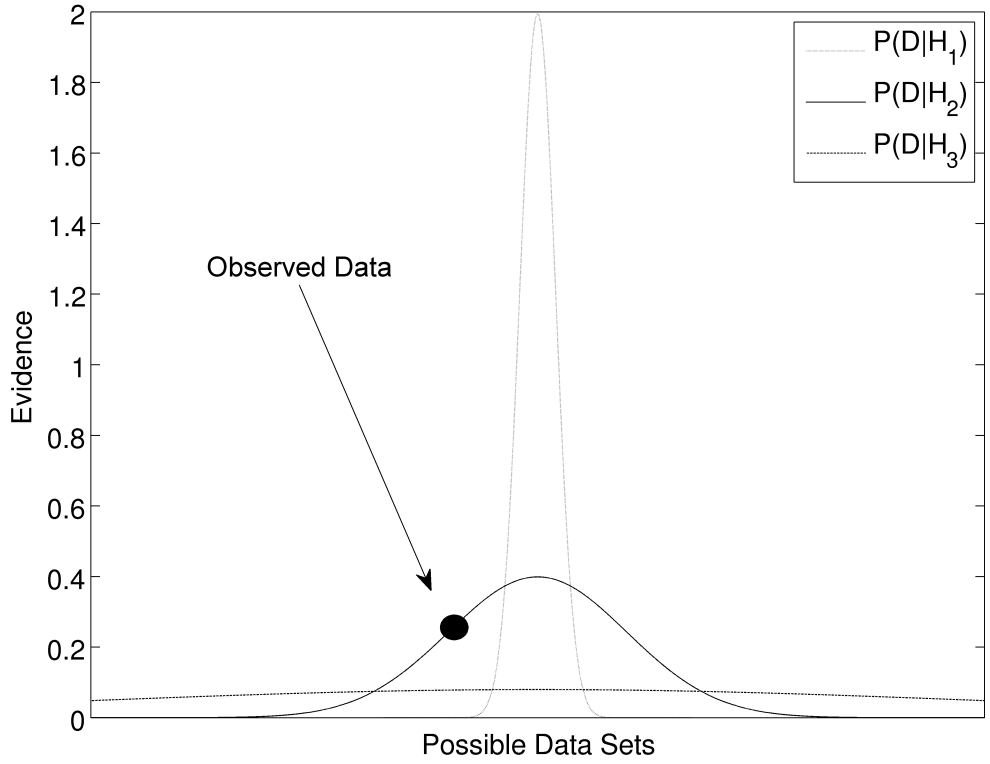


Figure 3.1: Adapted from MacKay [2003]. Observed data represents one of many possible data sets that could have been observed. Models that are too simple (H_1) fail to accurately fit the data. Models that are too complex (H_3) can explain too many alternatives with thinly spread probability mass. The correct model (H_2) fits the data parsimoniously, leading to a higher evidence and greater predictive power.

unsure whether the correct fit is a linear, quadratic or cubic model. The likelihood in each case is:

- Linear:

$$p(\,\boldsymbol{x} \mid \boldsymbol{t}, \boldsymbol{a}, M_{\text{linear}} I\,) = \prod_i \mathcal{N}\big(x_i; a_0 + a_1 t_i, \epsilon^2\big)\,. \tag{15}$$

- Quadratic:

$$p(\,\boldsymbol{x} \mid \boldsymbol{t}, \boldsymbol{b}, M_{\text{quadratic}}, I\,) = \prod_i \mathcal{N}\big(x_i; b_0 + a_1 t_i + b_2 t_i^2, \epsilon^2\big)\,. \tag{16}$$

- Cubic:

$$p(\,\boldsymbol{x} \mid \boldsymbol{t}, \boldsymbol{c}, M_{\text{cubic}}, I\,) = \prod_i \mathcal{N}\big(x_i; c_0 + c_1 t_i + c_2 t_i^2 + c_3 t_i^3, \epsilon^2\big)\,. \tag{17}$$

where $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$ are the parameters for each model, with the number of free parameters in each model being $k_{\text{linear}} = 2, k_{\text{quadratic}} = 3$ and $k_{\text{cubic}} = 4$. The maximum likelihood method can estimate the best-fit values for the parameters in each case but cannot determine which model is correct. This is easy to see since the models are 'nested'. The linear model is a subset of the quadratic model with $b_2 = 0$, and the quadratic model is a subset of the cubic model with $c_3 = 0$. Because we can reproduce the simpler models within the more complex ones the maximum likelihood for the more complex model is always guaranteed to be equal or greater than the maximum likelihood for the simpler one. Consider the generated quadratic data set shown in Figure 3.2. The MLE fit for all three models is shown, along with the value of the maximised likelihood and the marginal probability of the data for each model, having marginalised over the regression parameters. The cubic and quadratic fits are both a close match to the data. Using maximum likelihood to choose the model complexity over-fits the data, selecting the cubic model. Using the marginal likelihood reveals that the quadratic model is a better fit. In practice advocates of the MLE method would use a penalised likelihood for model selection, such as the Akaike Information Criterion (AIC) [Akaike, 1974] or the Bayesian Information Criterion (BIC) [Schwarz,

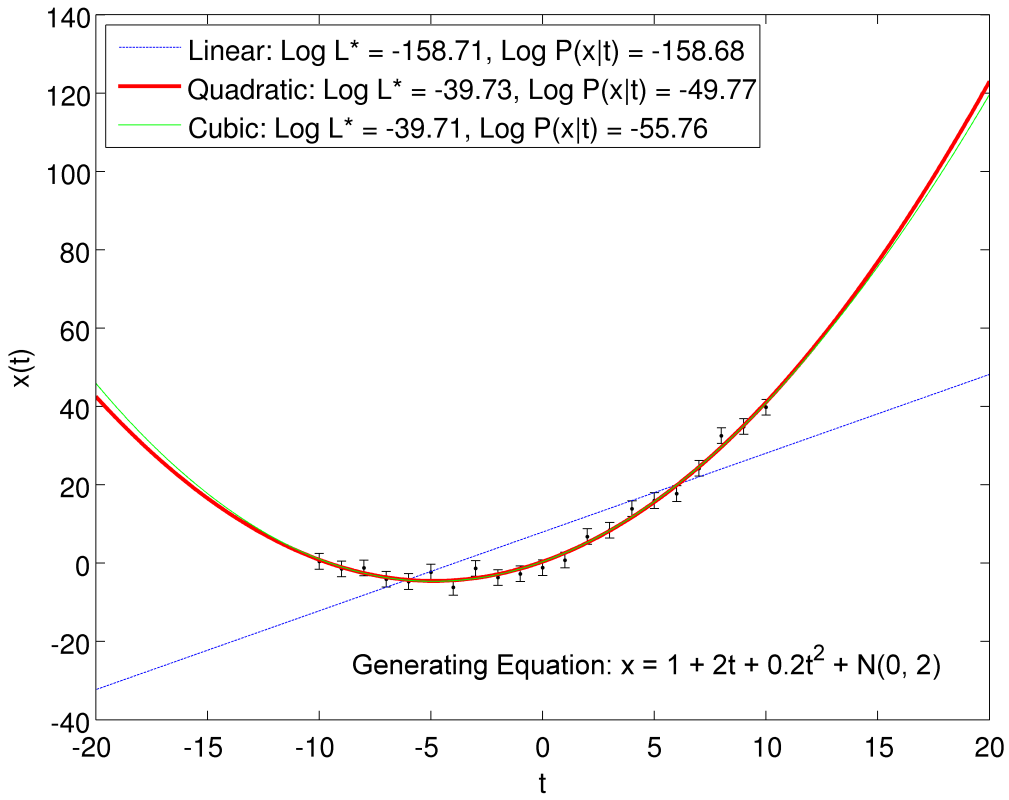


Figure 3.2: Three polynomial fits to a generated quadratic data set. Although the maximum likelihood ($L*$) is greatest for the cubic model, the marginal probability ($P(\, x \mid t \,)$) shows that the quadratic fit is the best.

1978]. In a simple case like this one these methods would detect the small change in likelihood between the quadratic and cubic models relative to the increase in the number of parameters and penalise the cubic model appropriately. However, for more complex models these criteria suffer from either not providing a probabilistic measure of the relative support for each hypothesis (AIC) or being an inexact approximation to the marginal likelihood (BIC). The marginal likelihood of each hypothesis remains the best way to choose the optimal model.

3.5 Case Study: Bayesian Linear Regression

Bayesian Linear Regression (BLR) provides an excellent case study to demonstrate all aspects of Bayesian inference and will prove useful in understanding the theory of Gaussian Processes in Chapter 4. Classical linear regression is one of the most widely used and understood methods of statistical analysis and a Bayesian treatment will illuminate the differences in approach.

Linear regression is the task of modelling data under the assumption that the output, $\boldsymbol{x}$, is linearly dependent on the (potentially multi-dimensional) input, $\mathbf{T}$. In the standard one-dimensional case this is the classic problem of identifying a 'best' straight line fit to some data, where the regression coefficients are simply the gradient and the intercept. Variation in the dependent variable around the straight line is usually assumed to take the form of indepedent and identically distributed Gaussian 'white noise'. The model is therefore expressed by the following likelihood, where $\boldsymbol{\beta}$ and ϵ are unknown parameters of the model:

$$p(\, \boldsymbol{x} \mid \boldsymbol{\beta}, \epsilon, I \,) = \mathcal{N}\big(\boldsymbol{x}; \mathbf{T}\boldsymbol{\beta}, \epsilon^2 \mathbf{E}_p\big) \tag{18}$$

where $\mathbf{E}_p$ represents the identity matrix of dimension p, and p is the length of $\boldsymbol{x}$. Note that from the start we are explicit about the probability model. In classical least squares regression the variation around the straight line is simply 'minimised' (in terms of square distance). The Bayesian model has two kinds of adjustable parameters: the regression coefficients, $\boldsymbol{\beta}$ and noise, ϵ (the intercept value can be appropri-

ately treated as an additional regression coefficient). Classical MLE linear regression obtains the same result as least squares regression (in cases where the noise model is Gaussian) by maximising equation (18) with respect to those parameters, obtaining an optimal set of regression coefficients,

$$\hat{\boldsymbol{\beta}} = (\mathbf{T}^\mathsf{T}\mathbf{T})^{-1}\mathbf{T}^\mathsf{T}\boldsymbol{x}. \tag{19}$$

In the Bayesian interpretation this idea is flawed in two ways. Firstly, as previously mentioned, to concentrate only on the likelihood is to confuse the inverse. Although the *data* may be most probable given a certain set of regression coefficients, it does not follow that these are the most probable coefficients. By including a prior over the coefficients we solve this problem. This prior needs to accurately reflect our knowledge about the likely values of $\boldsymbol{\beta}$ before we observe $\boldsymbol{x}$. See section 3.5.1 for a discussion on appropriate choice of prior. In this example we assume a Gaussian prior over the coefficients with zero mean and covariance, $\boldsymbol{\Sigma}_\beta$ and assume a known, fixed noise level, ϵ for simplicity.

$$p(\,\boldsymbol{\beta} \mid \boldsymbol{\Sigma}_\beta, I\,) = \mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, \boldsymbol{\Sigma}_\beta) \tag{20}$$

These are simply example choices and the prior used in any instance should reflect the particular problem at hand.

With this Gaussian prior we can use identities (A5) and (A7) to marginalise over the regression parameters and find the prior distribution for $\boldsymbol{x}$.

$$\begin{aligned} p(\,\boldsymbol{x} \mid I\,) &= \int p(\,\boldsymbol{x} \mid \boldsymbol{\beta}, I\,)\, p(\,\boldsymbol{\beta} \mid I\,)\, \mathrm{d}\boldsymbol{\beta} \\ &= \mathcal{N}(\boldsymbol{x}; \mathbf{0}_p, \boldsymbol{\Sigma}_D)\,, \end{aligned} \tag{21}$$

where $\mathbf{0}_p$ is a zero vector of length p, and with covariance matrix $\boldsymbol{\Sigma}_D$ given by,

$$\boldsymbol{\Sigma}_D = \epsilon^2\mathbf{E}_p + \mathbf{T}\boldsymbol{\Sigma}_\beta\mathbf{T}^\mathsf{T}. \tag{22}$$

The joint distribution for a training data set, $\boldsymbol{x}$ and a test data set, $\boldsymbol{x}_*$, of length q,

is also readily obtained through the same identities and is given as,

$$\begin{aligned} p\left(\begin{bmatrix} \boldsymbol{x}_* \\ \boldsymbol{x} \end{bmatrix} \,\middle|\, I \right) &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_* \\ \boldsymbol{x} \end{bmatrix} ; \begin{bmatrix} \mathbf{0}_q \\ \mathbf{0}_p \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_* & \boldsymbol{\Sigma}_{D*} \\ \boldsymbol{\Sigma}_{D*}^\mathsf{T} & \boldsymbol{\Sigma}_D \end{bmatrix} \right) \\ &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_* \\ \boldsymbol{x} \end{bmatrix} ; \mathbf{0}_{p+q}, \begin{bmatrix} \mathbf{T}_* \\ \mathbf{T} \end{bmatrix} \boldsymbol{\Sigma}_\beta \begin{bmatrix} \mathbf{T}_* \\ \mathbf{T} \end{bmatrix}^\mathsf{T} + \epsilon^2 \mathbf{E}_{p+q} \right), \end{aligned} \tag{23}$$

with additional covariance matrix components given by,

$$\begin{aligned} \boldsymbol{\Sigma}_* &= \epsilon^2 \mathbf{E}_q + \mathbf{T}_* \boldsymbol{\Sigma}_\beta \mathbf{T}_*{}^\mathsf{T}, \\ \boldsymbol{\Sigma}_{D*} &= \mathbf{T} \boldsymbol{\Sigma}_\beta \mathbf{T}_*{}^\mathsf{T}. \end{aligned} \tag{24}$$

At this point we have a multi-variate Gaussian distribution over the observables, $\boldsymbol{x}$ and $\boldsymbol{x}_*$, with a zero mean vector and a covariance matrix that is a function of the regressors, $\mathbf{T}$ and $\mathbf{T}_*$. It is important to recognise this, since we will return to this idea when discussing Gaussian processes in the next chapter.

If we are interested in the value of the regression coefficients, as opposed to predicting observable data, we can also calculate the posterior distribution of these parameters through identity (A7) and using the Woodbury matrix identity [Petersen and Pedersen, 2008] to simplify the covariance matrix, giving the posterior as,

$$p(\, \boldsymbol{\beta} \mid \boldsymbol{x}, \mathbf{T}, \boldsymbol{\Sigma}_\beta, I \,) = \mathcal{N}\big(\boldsymbol{\beta}; \epsilon^{-2} \boldsymbol{\Sigma}'_\beta \mathbf{T}^\mathsf{T} \boldsymbol{x}, \boldsymbol{\Sigma}'_\beta\big)\,, \tag{25}$$

where the updated covariance matrix, $\boldsymbol{\Sigma}'_\beta$ is given by,

$$\boldsymbol{\Sigma}'_\beta = (\epsilon^{-2} \mathbf{T}^\mathsf{T} \mathbf{T} + \boldsymbol{\Sigma}_\beta^{-1})^{-1}. \tag{26}$$

Note that we recover the MLE for $\boldsymbol{\beta}$ in the limit as the prior distribution over $\boldsymbol{\beta}$ becomes diffuse (as the covariance becomes large).

Having acquired the posterior we might feel inclined to locate the maximum value to determine an 'optimal' choice for the coefficients in similar fashion to the MLE. While this would certainly be preferable to the MLE since it properly incorporates our prior information it would ignore the true power of the Bayesian method. We should ask ourselves 'why do we want to find the regression coefficients?' If the

parameters themselves are what we want then we should report the posterior as fully as possible, typically through the mean and variance of the distribution, although in some cases further or more representative moments of the distribution would be required. Thus we will communicate as much information as we have. Otherwise, if we simply choose a single 'optimum value' we risk selecting a point estimate that is unrepresentative of the probability distribution of the regression coefficients and produces inaccurate, biased or overly-confident predictions.

If, however, we aim to predict the values of $\boldsymbol{x}_*$ for hypothetical new values of $\mathbf{T}_*$, based on observing $\boldsymbol{x}$, then we can make these predictions without ever having to choose a particular value of $\boldsymbol{\beta}$. By marginalising over our knowledge of $\boldsymbol{\beta}$, or equivalently by applying identities (A2) and (A4) along with equation (23), we obtain an optimal prediction with an honest representation of the uncertainty.

$$p(\, \boldsymbol{x}_* \mid \mathbf{T}_*, \boldsymbol{x}, \mathbf{T}, \epsilon, \boldsymbol{\Sigma}_\beta, I \,) = \mathcal{N}\big(\boldsymbol{x}_*; \boldsymbol{\Sigma}_{*D}\boldsymbol{\Sigma}_D^{-1}\boldsymbol{x}, \boldsymbol{\Sigma}_* - \boldsymbol{\Sigma}_{*D}\boldsymbol{\Sigma}_D^{-1}\boldsymbol{\Sigma}_{*D}^{\mathsf{T}}\big) \tag{27}$$

3.5.1 Choice of prior

At first we might insist that we are entirely ignorant regarding the likely values of $\boldsymbol{\beta}$ and argue for a uniformly flat prior, giving equal weight to all possible values from $-\infty$ to $+\infty$. In this case the posterior would be exactly proportional to the likelihood and their would be no confusion of the inverse. Many Bayesian analyses do in fact claim to use this prior to express their initial ignorance. However, it is not difficult to see that this idea is flawed. To begin with we ought to question whether the claim that this prior is being used is in fact true. In most cases *it is not*. Using this prior is equivalent to arguing that the posterior is proportional to the likelihood over all space, differing only by a normalising constant. But what is that constant? It will be the integral of the likelihood over all space, so as to assure the posterior integrates to unity. But there is no guarantee that such an integral converges at all. Because the likelihood is the probability of the *data* given a particular parameter value it must sum to unity in the data space, but need not in general sum to unity (or to any finite

value) in the parameter space. That is,

$$\begin{aligned} \sum_{\boldsymbol{x}} L(\boldsymbol{\beta}) &\equiv \sum_{\boldsymbol{x}} P(\,\boldsymbol{x} \mid \boldsymbol{\beta}, I\,) \equiv 1 \\ \sum_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) &\equiv \sum_{\boldsymbol{\beta}} P(\,\boldsymbol{x} \mid \boldsymbol{\beta}, I\,) \neq \text{Constant} \quad \text{(In general)}. \end{aligned} \tag{28}$$

Furthermore, in most instances the likelihood is only evaluated over a subset of the possible (often infinite) parameter space. There is no guarantee that huge areas of large likelihood values do not lie outside this bound. In practice, most instances where the flat uniform prior is used technically result in some hard bounds being placed on that prior. This is clearly a strong rejection of ignorance, effectively stating 'the parameter can be this high but no higher'. True use of the *improper* uniform prior over all space would in effect be arguing that we can know nothing about the posterior until we have explored the likelihood over the infinite parameter space, leading to an absurd situation in which inference is not possible.

The Bayesian approach is to accept that we do hold (generally weak) information about the possible parameter values. We ought to seek to convey this information through an appropriate *proper* prior — that is one that represents a genuine probability distribution that can be summed to unity. We might find it difficult to translate the vague knowledge we hold (e.g. previous experience of similar data types, judgements about what sort of relationships seem sensible etc) into concrete probability distributions, but this is a flaw in our internal information storage, not in probability theory. It can be argued that this may introduce an unacceptable element of subjectivity into the analysis. The counter to this argument is first to acknowledge the subjectivity. If two agents have genuinely different beliefs before they observe data they will continue to hold different beliefs after updating to the posterior. If both update according to the rules of probability theory then with sufficient data their beliefs should converge towards each other as the posterior becomes less and less dependent on the prior. Therefore the subjective disagreement of different agents is a sign that insufficient data has been observed to settle the dispute.

There are some theoretical ideas regarding appropriate choice of priors through

maximum entropy methods [Brewer and Francis, 2009]. Principles of invariance may be used to construct an 'objective prior', such as the Jeffreys prior [Jeffreys, 1946] but these can lead to improper prior distributions that cannot represent true probabilities. While there is fertile ground for further research into the construction of objective priors that can be agreed on by all, this is only of concern in the subset of problems where the prior has a strong effect on the eventual outcome of the analysis. When there is sufficient data to overcome the influence of all but the most pathological prior distributions, a *reasonable* prior is one that is sufficient to cover a wide range of the parameter space around the expected values, and which complements basic characteristics of the parameter space.

The Gaussian distribution is, in many cases, an ideal choice as it maximises the entropy for a given mean and variance and often provides useful analytic tractability. As discussed above the exact form of the prior is important only in cases where there are insufficient data. The Gaussian distribution is the optimal choice when we have some idea what the value of the parameter should be and some idea of the uncertainty in that estimate, especially when our prior information also includes the constraint of limited computational time. We should aim to make the best possible inference within these constraints.

3.5.2 Towards Gaussian processes

Linear regression posits that that the observables are noisy observations of a latent linear function of the regressors. Clearly a great many other functional relationships are possible, and in the next chapter Gaussian processes are introduced to provide a more flexible framework for inferring those relationship. Gaussian processes are introduced by considering the joint Gaussian distribution of many variables. As will eventually become clear, BLR with a Gaussian prior on the regression coefficients is a special example of a Gaussian process, using a particular covariance between the elements of $\boldsymbol{x}$ derived from the respective elements of $\mathbf{T}$.

An alternative framework for inferring more complicated functional relation-

ships is non-linear regression. Non-linear regression can be performed in the same manner as linear regression, by exchanging the regressors, $\mathbf{T}$ for a set of non-linear basis functions of those regressors, $\phi(\mathbf{T})$. For example, quadratic regression is performed with the basis function, $\phi(\mathbf{T}) : T_{ij} \rightarrow T_{ij}^2$. This is an alternative route to understanding Gaussian processes than the route taken in this thesis, arriving at Gaussian processes as the limiting case of regression over infinitely many basis functions, the form of which determines the structure of the resulting Gaussian process. This is explored by Rasmussen and Williams [2006] (section 2.1.2). For the problems modelled in this thesis the approach taken in the next chapter is more natural than extending regression, but both approaches conclude in the same mathematical formulation.

Chapter 4

Gaussian Processes

4.1 Introduction

Gaussian processes (GPs) [Stein, 1999, MacKay, 2003, Rasmussen and Williams, 2006] are a powerful and flexible framework for performing Bayesian inference over functions. This chapter describes GPs by initially discussing the multi-variate Gaussian distribution, using a simple bi-variate case study to illustrate the basic ideas. The fundamental structure and equations of a Gaussian process model are then presented before a discussion of the role of covariance functions and the associated hyper-parameters in creating an appropriate model, and how covariance functions can be combined to create more complex models.

4.2 Functions

A *function* specifies a mapping from a domain, the *input values*, to a codomain, the *output values*. We can specify a finite number of input values as the components of a vector. For the sake of future clarity of nomenclature let us assume that the input values represent a time index. A vector of input values may then be denoted as $\mathbf{t}$, where an individual component t_i represents the ith time index. With each input value there is an associated output value — we exclude functions that have multiple outputs for a single input. We will denote the output value associated with input value t_i as $x(t_i)$ and the set of output values as $x(\mathbf{t})$.

4.3 Gaussian Distribution to Gaussian Processes

As Rasmussen and Williams [2006] demonstrate, GPs can be approached in several ways. For the purposes of this thesis the 'function-space' approach is most intuitive. This directly views GPs as probability distributions over function-valued random variables. We use the following notation, similar to the Gaussian distribution. If the random variable X has a GP distribution then,

$$p(\, X = x(t) \mid I \,) \equiv p(\, x(t) \mid I \,) = \mathcal{G}P(x(t); m(t), k(t, t')) \,, \tag{29}$$

where $m(t)$ is the *mean function* and $k(t, t')$ represents the *covariance kernel*. We will see how these concepts emerge through this chapter.

To see how GPs can be used to perform inference over functions, begin by examining the finite multi-variate Gaussian distribution. A simple example in two-dimensions is shown in Figure 4.1. The bi-variate Gaussian distribution specifies a joint probability distribution of two potentially co-varying variables, given in three different notations by the equalities below:

$$\begin{aligned} P(\, x, y \mid I \,) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\quad \times \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{(\sigma_x\sigma_y)}\right)\right), \\ &= \frac{1}{(2\pi)^{N/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \\ &\equiv \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,, \end{aligned} \tag{30}$$

where ρ represents the strength of the correlation between x and y. The second notation presents the probability in matrix form with a change of variables, $\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$. This matrix form will remain the same as the distribution is extended to include more variables. The third form uses the notation used throughout this thesis for any Gaussian distribution over a finite number of variables.

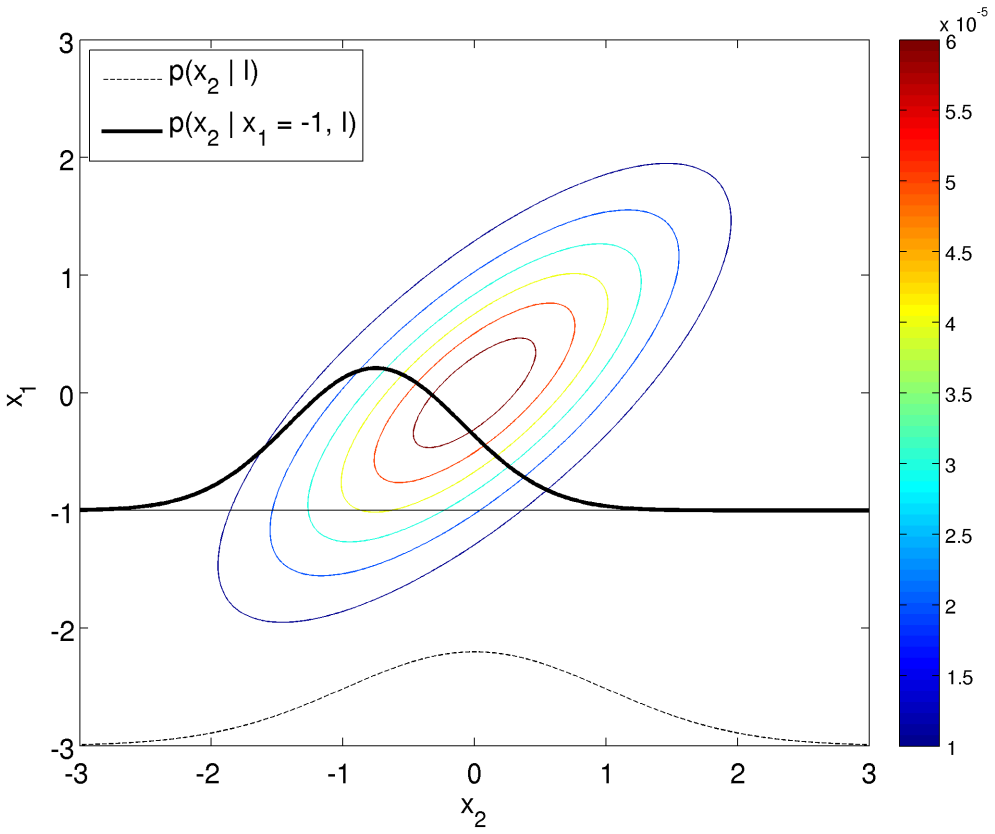


Figure 4.1: Bi-variate Gaussian distribution. The contour plot shows the joint probability density function of two jointly Gaussian random variables, with correlation $\rho = 0.75$. The dashed line shows the marginal distribution of x_2, while the heavy line shows the conditional distribution of x_2 after making the observation: $x_1 = -1$.

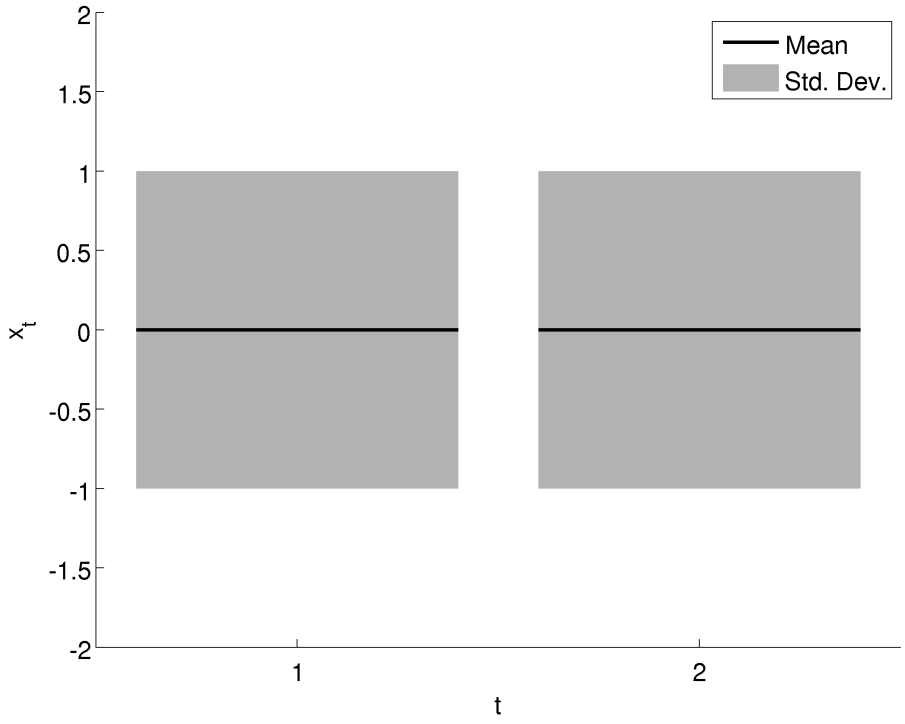
The bi-variate case is illuminating because it demonstrates the key properties of the multi-variate Gaussian distribution. The distribution of one of the variables is individually Gaussian without the other. This is referred to as being *marginally Gaussian*

$$P(\, y \mid I \,) = \int P(\, x, y \mid I \,)\, \mathrm{d}x = \mathcal{N}\big(y; \mu_y, \sigma_y^2\big) \tag{31}$$

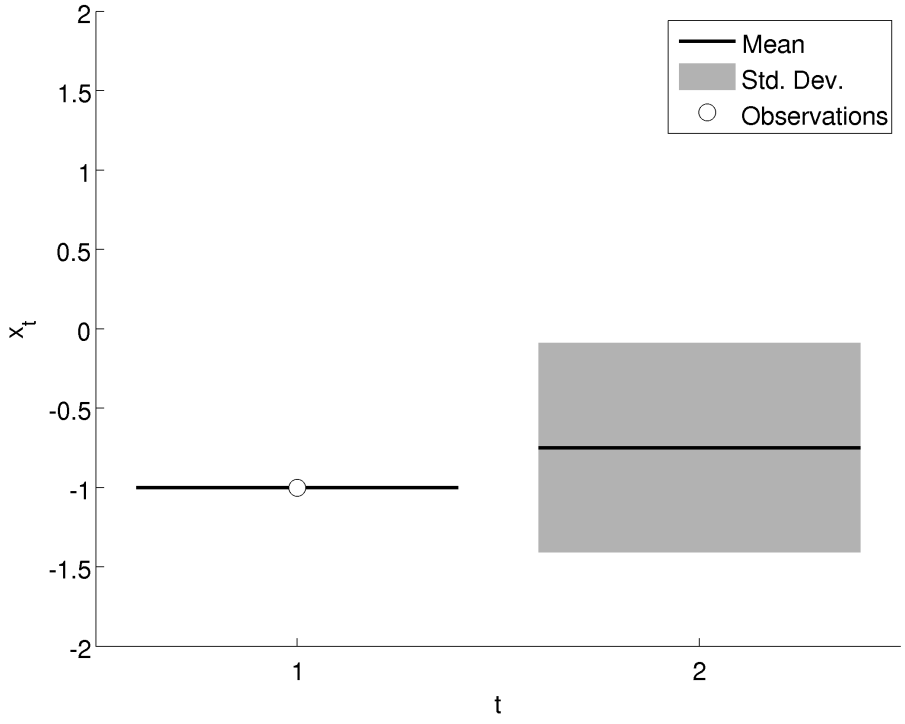
Furthermore the distribution of y remains Gaussian when we learn the value of x. This is referred to as being *conditionally Gaussian*

$$P(\, y \mid x, I \,) = \frac{P(\, y, x \mid I \,)}{P(\, x \mid I \,)} = \mathcal{N}\big(y; \mu_y + \rho(x - \mu_x), \sigma_y^2 - \rho^2/\sigma_x^2\big) \tag{32}$$

The extensions of equations (31) and (32) to more variables can be found in the Appendix A, equations (A2), (A3) and (A4). Equation (32) is especially important. Note that by observing x the variance on y is potentially reduced (and the expectation adjusted) when $\rho \neq 0$. Thus if there exists a covariance between x and y, observing x informs us about y. This can be seen graphically in Figure 4.2. We can extend this idea to more dimensions, representing our knowledge of the variables similarly to Figure 4.2. In a general many-variable system the relationships between variables can be arbitrarily strong or weak (technically a covariance matrix must be positive semi-definite — this ensures that the covariances between variable pairs are not contradictory). To move towards the standard framework of Gaussian processes the most useful examples are those where variables with similar indices have strong correlations and those with dissimilar indices have weak correlations. In this case the closer an unknown variable is to an observed variable the lower its uncertainty. This is demonstrated in Figure 4.3. Plot (a) shows a prior distribution over ten variables before making any observations, showing the marginal mean and standard deviation of each variable; each variable is identically distributed. Plot (b) shows the posterior after making three observations on variables x_3, x_5 and x_8. Plots (c) and (d) represents the generalisation to a distribution of a infinite continuum of variables, with the expectation forming a smooth fit to the observations and the uncertainty increasing smoothly with distance to the closest observation. This is done by extending the



(a)



(b)

Figure 4.2: Effect of making an observation in a two variable system: In plot (a) both variables (x_1 and x_2) are equally uncertain with the same distribution as Figure 4.1. In plot (b) is shown the effect of making a perfect observation of $x_1 = -1$. The observed variable is now known. The expectation of x_2 is adjusted downwards and the standard deviation reduced according to equations (31) and (32).

formalism of the multi-variate Gaussian distribution to cover an infinite number of variates, as explained below.

Infinite variable distribution

Equation 30 provides the general matrix form for the probability density in a multi-variate Gaussian distribution. The task now is to extend this from a finite collection of variables to a function-valued variable, which represents an infinite number of variates. Let us assume we have observed a function $x(t)$ at a finite set of sampling points, $\boldsymbol{t}$, observing function values $x(\boldsymbol{t})$. Rasmussen and Williams [2006] give a definition for a Gaussian Process as:

Definition: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

This implies that if X is a function-valued random variable that has a GP distribution, any finite subset of X has a Gaussian distribution, therefore

$$\begin{aligned} p(\, X = x(t) \mid I \,) \equiv p(\, x(t) \mid I \,) &= \mathcal{G}P(x(t); m(t), k(t,t')) \\ \implies p(\, x(\boldsymbol{t}) \mid I \,) &= \mathcal{N}(x(\boldsymbol{t}); m(\boldsymbol{t}), k(\boldsymbol{t},\boldsymbol{t}))\,, \end{aligned} \tag{33}$$

where $k(\boldsymbol{t},\boldsymbol{t})$ represents the matrix evaluation of $k(t,t')$ for all possible pairs of input values from the vector $\boldsymbol{t}$.

In this case the full collection of random variables are all the values of $x(t)$, which is equivalent to a vector of infinite length, with one element for every potential input value. The finite collection are the sampled values, $x(\boldsymbol{t})$. As per the above definition we can model the entire function by assuming that any finite set of samples obeys a multi-variate Gaussian distribution. The mean, $m(t)$ is simply a function that describes the expected value of $x(t)$ before we make observations, and the mean for a finite set is simply $m(\boldsymbol{t})$. The *covariance kernel* is defined over all potential pairs of inputs, and represents the covariance of function output values in terms of

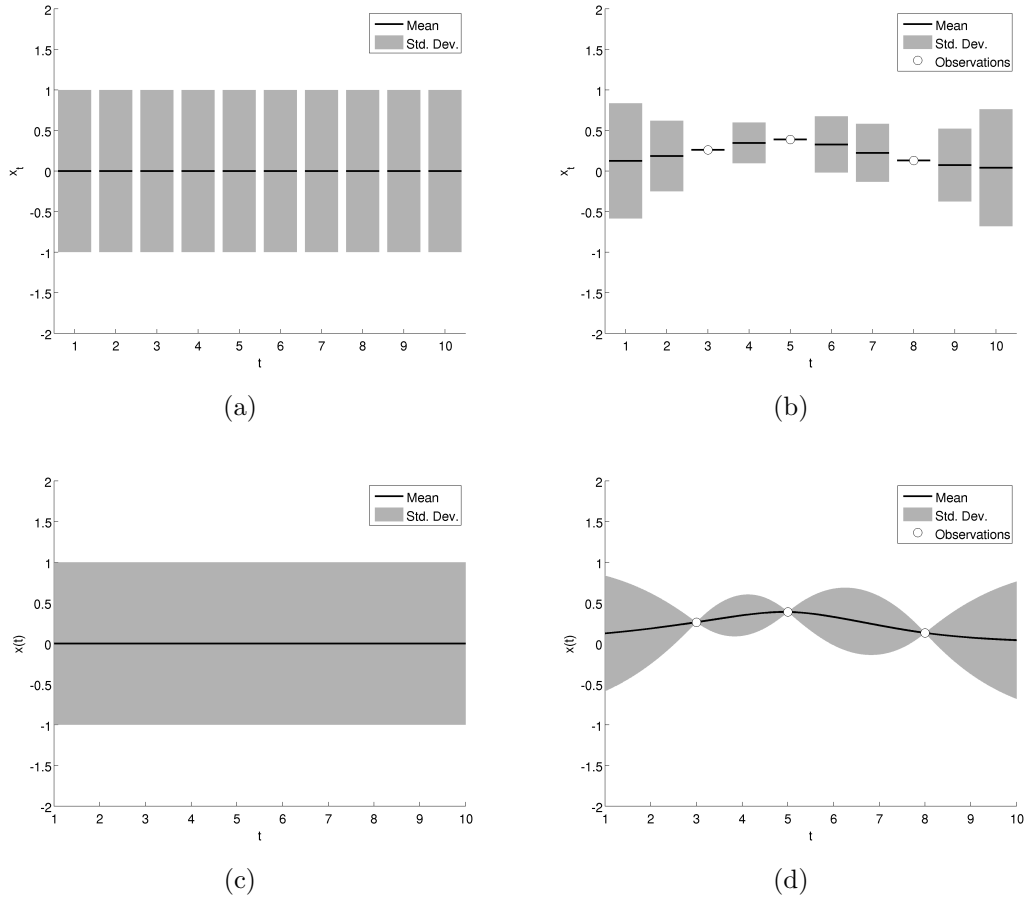


Figure 4.3: Extending the multi-variate Gaussian distribution to an infinite number of variables. In panels (a) and (b) a 10-variate Gaussian distribution is represented similarly to Figure 4.2. Panel (a) shows the distribution before any observations are made, panel (b) shows the distribution after three observations at made at $\boldsymbol{t} = [3, 5, 8]$. Each variable is most strongly correlated with those that have a similar value of t, therefore the standard deviation is lowest near the observations and highest. Panels (c) and (d) show the analogue of this distribution extended to an infinite number of variables - the real numbers between 1 and 10. Now the standard deviation grows smoothly with distance from the closest observation.

their inputs,

$$\mathrm{Cov}(x(t_i), x(t_j)) = \Sigma_{ij} = k(t_i, t_j). \tag{34}$$

The covariance kernel must be a positive semi-definite function, meaning that any set of inputs must produce a covariance matrix, $\mathbf{\Sigma} = k(\boldsymbol{t}, \boldsymbol{t})$, with elements given by equation (34), which is positive semi-definite (i.e. one that has no negative eigenvalues). The distribution of different finite sets must be consistent. Colloquially this means that if one function value is highly correlated with a second, which is in turn highly correlated with a third, there must also exist a significant correlation between the first and third values. The particular functional form of the covariance function is subject to restrictions to ensure the positive semi-definite requirement for the covariance matrix. In addition there will be further conditions imposed by choice — see section 4.4 for more details.

The mean and covariance functions define the types of outputs we expect to see from the GP distribution. The mean specifies what we expect the function output values to be, prior to making any observations. In simple problems a popular choice is simply to set $m(t) = 0$. In more complex problems the mean may naturally be specified by reference to some symmetry or invariance in the problem at hand. For example, in the case of a bird flying between a release point and its home loft the mean will be specified *a priori* as the straight line path between those two locations, thereby reflecting a key symmetry in the particular problem.

The covariance function determines variation of output around the mean function with changing input for the functions we aim to model. An appropriate choice of covariance function can specify that the function is smooth and slowly varying, or rapidly varying and highly disordered. The covariance can also incorporate more detailed knowledge of the function, such as potential periodicity or the existence of 'change-points' — where the function values exhibit little or no relation across a boundary. The next section will discuss the form of the covariance functions used in this thesis. See Rasmussen and Williams [2006], chapter 4 for further discussion of

covariance forms beyond the scope of this thesis.

So far we have specified a *prior* distribution over the function before we make any observations. Now assume that we have observations, $x(\boldsymbol{t}_D)$ and we are interested in making predictions about the value of the function, $x(\boldsymbol{t}_*)$ at new inputs $\boldsymbol{t}_*$. Then by equations (A2) and (A4) (Appendix A), the *posterior* distribution of these values is given as

$$p(\, x(\boldsymbol{t}_*) \mid x(\boldsymbol{t}_D), I\,) = \mathcal{N}(x(\boldsymbol{t}_*); \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)\,, \tag{35}$$

with updated mean and covariance matrices:

$$\boldsymbol{\mu}_* = m(\mathbf{t}_*) + k(\mathbf{t}_*, \mathbf{t}_D)k(\mathbf{t}_D, \mathbf{t}_D)^{-1}(x(\boldsymbol{t}_D) - m(\boldsymbol{t}_D)) \tag{36}$$

$$\boldsymbol{\Sigma}_* = k(\mathbf{t}_*, \mathbf{t}_*) - k(\mathbf{t}_*, \mathbf{t}_D)k(\mathbf{t}_D, \mathbf{t}_D)^{-1}k(\mathbf{t}_D, \mathbf{t}_*). \tag{37}$$

4.4 Covariance Functions

The covariance function expresses the covariance or correlation of pairs of output values as a function of their respective input values. In the most widely used examples this is simply a function of the separation between the input values,

$$\mathrm{Cov}(t_i, t_j) = k(\Delta t \equiv |t_i - t_j|). \tag{38}$$

These are known as *stationary* covariances since the form is independent of the absolute input values and dependent only on the difference Δt. In this thesis I will restrict my analysis to these cases. In particular I will examine a range of simple covariances that are monotonically decreasing functions of Δt. This implies that similar inputs should result in similar outputs, imposing a degree of *smoothness* on the output values. The standard example of this type is the *squared exponential* covariance, $k_{SE}(\Delta t)$,

$$k_{SE}(\Delta t) \equiv \lambda^2 \exp\left(\frac{-\Delta t^2}{2\sigma^2}\right). \tag{39}$$

This functional form includes two adjustable 'hyper-parameters', σ and λ, which I will term the *input scale* and *output scale* respectively. Their role is discussed below.

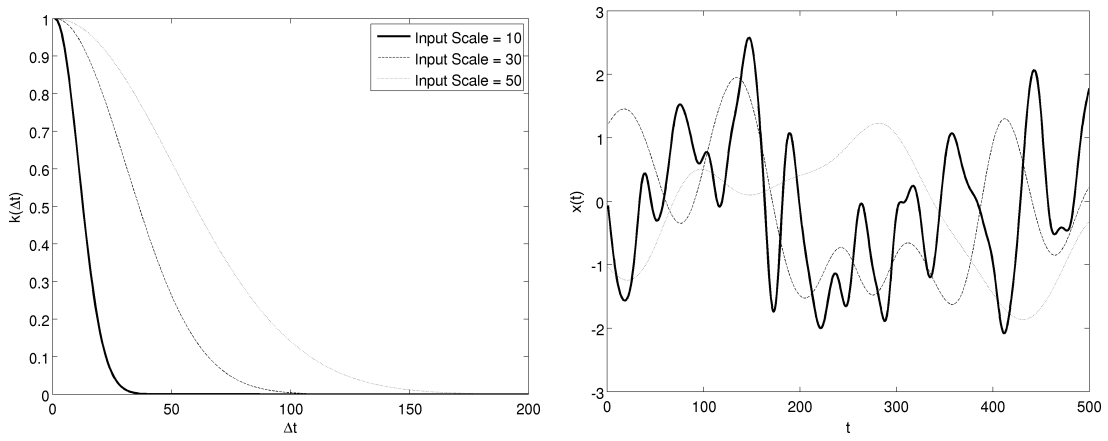


Figure 4.4: Examples of the Squared-Exponential covariance with varying input scale, σ. The left panel shows the covariance as a function of input separation. the right panel shows a random sample from a GP using each covariance. For all cases $\lambda = 1$. As the input scale is increased the output becomes smoother and varies more slowly.

4.4.1 Hyper-parameters

The adjustable parameters in (39) alter the distribution of function output values from the GP. As the names given to them suggest, the input scale is a scaling in the input space and the output scale a scaling in the output space.

Larger input scales increase the range of correlations as a function of the input value separation, leading to greater coupling between distant output values. This makes the resulting functions smoother and more slowly varying. Smaller input scales conversely lead to more rapidly varying functions where output values become uncorrelated with smaller differences in inputs. This is demonstrated in Figure 4.4. In the left hand panel the squared exponential covariance function is evaluated over a range of input separations for a variety of different input scales. The adjacent panel shows random samples from a GP using the corresponding covariance function. The smaller input scales lead to rapidly varying samples compared to the smoothly varying samples from the longer input scales. The output scale simply acts as a rescaling factor for the entirety of the output function. This is equivalent to a rescaling of the vertical axes in Figure 4.4, such that a doubling of the output scale will double the values of both the covariance functions and the corresponding output samples.

Typically we are unsure about the values of these hyper-parameters *a priori* and are restricted to specifying a prior that accurately represents that uncertainty. The data of observed function values will be informative of the correct values and will reduce our uncertainty through the use of Bayes' rule. In the light of uncertainty we will marginalise over these hyper-parameters by integration, typically relying on Monte Carlo techniques where an analytic answer is not available (see Appendix B).

4.4.2 Other covariance functions

The squared-exponential function is the most commonly employed covariance kernel in applications of GPs. However, it has certain properties that potentially make its use problematic. The squared-exponential function fails to accurately represent the structure of variation in certain types of data, particularly the output of real physical systems as opposed to mathematically defined test data. There are good physical reasons to doubt the suitability of the squared-exponential function in the particular system dealt with in this thesis. Samples from a GP with a squared-exponential covariance have the property of *infinite mean-square differentiability* (see Rasmussen and Williams [2006]). While mean-square differentiability of the process does not imply differentiability of samples from that process, this property does imply very high very-local smoothness in the samples. This is a perfectly valid assumption for most mathematically generated data or, for example, a system composed of particles moving in a continuous field, where force varies continuously with position, but it is likely untenable for an autonomous agent such as a bird which is capable of self-accelerating at arbitrarily chosen instances. We would expect real paths (excluding observation noise) to be once differentiable, corresponding to continuity in velocity. Further differentials of the path with respect to time should result in discontinuities, reflecting instantaneous changes in acceleration by the bird.

Stein [1999] argues that the more general Matérn class of covariance functions are better suited to modelling physical systems. Of course, adhering to the principles of Chapter 3, the best test for the most appropriate covariance function is to calculate

the evidence for each from real data. This is supported by my own findings which indicate a greater evidence for models based on Matérn functions than for models using the squared-exponential (see Chapter 5, section 5.5).

The Matérn class obeys a general form with an adjustable *differentiability parameter*, ν (form adapted from Rasmussen and Williams [2006]),

$$k_\nu(\Delta t) \equiv \lambda^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \left| \frac{\Delta t}{\sigma} \right| \right)^\nu \mathfrak{K}_\nu \left(\sqrt{2\nu} \left| \frac{\Delta t}{\sigma} \right| \right), \tag{40}$$

where $\mathfrak{K}_\nu$ is a modified Bessel function [Abramowitz and Stegun, 1965]. As with the earlier definition of the squared-exponential covariance function, σ and λ represent the input and output scale hyper-parameters. The differentiability parameter, ν, adjusts the very localised smoothness of samples from the GP with this covariance (whereas the input-scale, σ, determines the smoothness of samples over a wider range). Samples generated by a GP with a Matérn covariance will be n times mean-square differentiable if $\nu > n$. The greater the value of ν the smoother the output will be over very small scales. The squared-exponential emerges as a special case as ν tends to infinity.

The general Matérn function becomes significantly simpler in the case where $\nu = n + 1/2$. Of these cases Rasmussen and Williams [2006] state that the most interesting examples are $\nu = 3/2$ and $\nu = 5/2$, arguing that for $\nu = 7/2$ or higher it is difficult to distinguish between the Matérn function and the squared-exponential. The output from a GP using $\nu = 1/2$ is highly disordered and is also inappropriate for the type of animal movement data analysed in this thesis. See Figure 4.5 for a comparison of the Matérn function with varying ν and resulting outputs. The simplified forms for the Matérn functions used in this thesis are given by,

$$k_{\nu=3/2}(\Delta t) \equiv \lambda^2 \left(1 + \frac{\sqrt{3}\Delta t}{\sigma} \right) \exp\left(-\frac{\sqrt{3}\Delta t}{\sigma} \right), \tag{41}$$

$$k_{\nu=5/2}(\Delta t) \equiv \lambda^2 \left(1 + \frac{\sqrt{5}\Delta t}{\sigma} + \frac{5\Delta t^2}{3\sigma^2} \right) \exp\left(-\frac{\sqrt{5}\Delta t}{\sigma} \right). \tag{42}$$

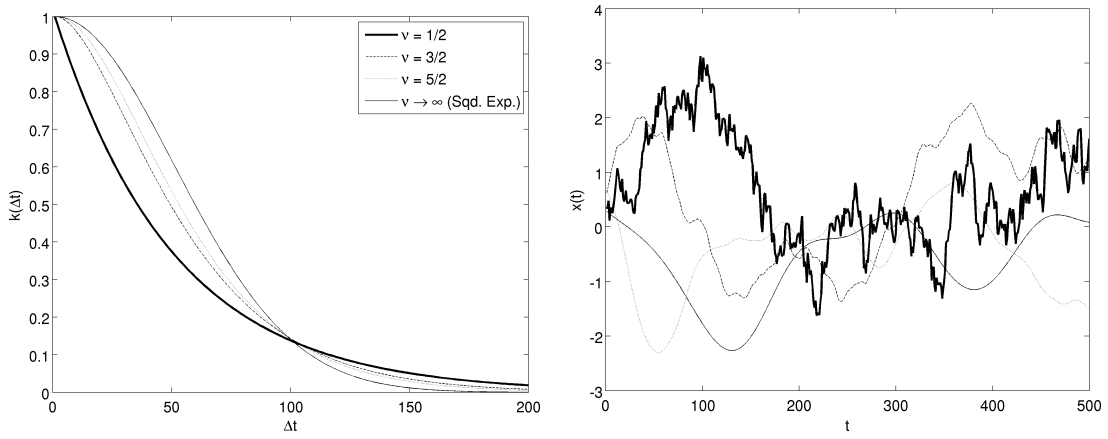


Figure 4.5: Examples of the Matérn covariance with varying differentiability parameter, ν. (A) shows the covariance as a function of input separation. (B) shows a random sample from a GP using each covariance. For all cases $\sigma = 50$, $\lambda = 1$. As ν is increased the covariance function becomes wider at small separations, leading to locally smoother outputs. Brownian motion corresponds to the special case $\nu = 1/2$. As ν tends to infinity we recover the squared-exponential function.

4.4.3 Observation noise and combining multiple covariance functions

It is frequently the case that we want to model a particular problem to account for multiple possible features in the data. For example, we may suspect that a data series has an overall trend superimposed on a periodic pattern such as seasonal variation. We may suspect that our observations of some process are imperfect and subject to 'observation noise' that constitutes an addition to the underlying pattern. In these cases we need to use multiple covariance functions in our model.

To illustrate, consider a model including noisy observations. Imagine we aim to model a process, $x(t)$, using a GP. However, we can only observe a noisy set of data $y(x)$, consisting of $x(t)$ and an additive zero-mean process with variance ϵ^2. The distribution over $y(x)$ is hence,

$$p(\, y(x(t_i)) \mid x(t_i), I \,) = \mathcal{N}\big(y(x(t_i)); x(t_i), \epsilon^2\big) \tag{43}$$

Or in vector form and assuming the noise process is uncorrelated through time,

$$p(\, y(x(\boldsymbol{t})) \mid x(\boldsymbol{t}), I \,) = \mathcal{N}\big(y(x(\boldsymbol{t})); x(\boldsymbol{t}), \epsilon^2 \mathbf{E}_p\big)\,, \tag{44}$$

where $\mathbf{E}_p$ is the $p \times p$ identity matrix and p is the length of the vector $\boldsymbol{t}$. In addition we have a GP for the process $x(t)$,

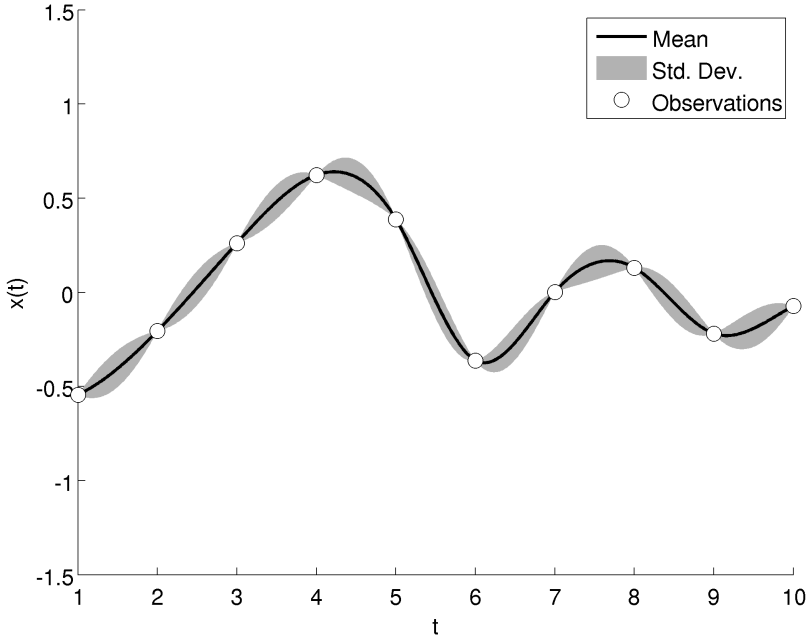
$$p(\, x(t) \mid I \,) = \mathcal{G}P(x(t); m(t), k(t, t')) \,. \tag{45}$$

We may combine equations (45) and (44) through identities (A5) and (A7) to obtain the full distribution for $y(\boldsymbol{t})$, marginalised over the unseen value of $x(t)$:

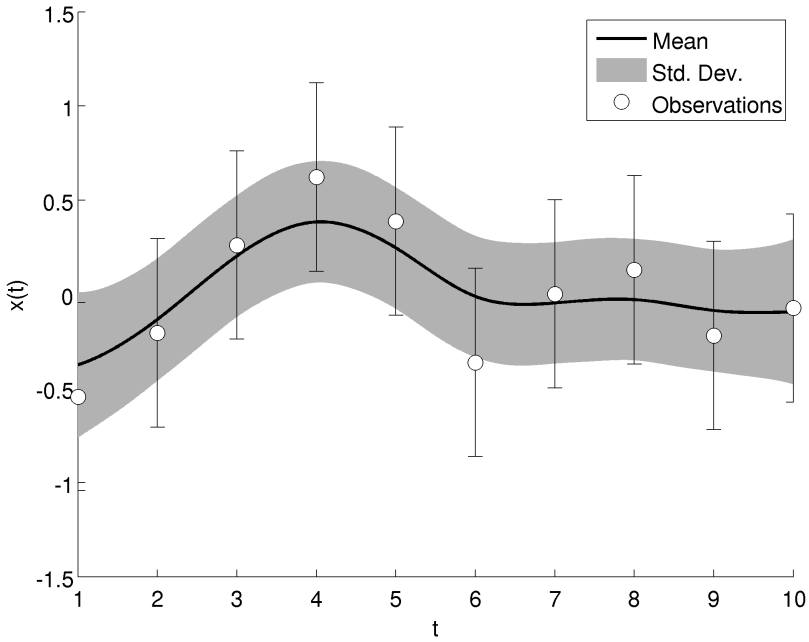
$$p(\, y(\boldsymbol{t}) \mid \boldsymbol{t}, I \,) = \mathcal{N}\big(y(\boldsymbol{t}); m(\boldsymbol{t}), k(\boldsymbol{t}, \boldsymbol{t}) + \epsilon^2 \mathbf{E}_p\big) \tag{46}$$

Figure 4.6 shows how the same underlying GP model fits a given data set with noiseless observations and with observation noise included. In cases where the observation noise is not known in advance, multiple interpretations of the same data may be possible. As the figure shows, observed variation can be explained either as variation in the latent function (plot (a)) or as observation noise imposed on a smooth latent function (plot (b)). The result derived above is valid beyond the case of observation noise. By replacing the identity matrix in equation (44) with any valid covariance we can show that the sum of multiple covariance matrices is also a valid covariance. We may easily use this to combine uncertainty. If, for example, we believe a data set exhibits variation on two scales we can simply use a covariance function comprised of two standard squared-exponential or Matérn functions with different hyper-parameters. Alternatively, if we expect the data to take the form of a periodic signal combined with a linear or polynomial drift we could use the addition of a periodic covariance with a covariance appropriate for linear or polynomial regression (see Rasmussen and Williams [2006], Chapter 4 for examples). This would be the case, for example, in modelling climate variables that may increase or decrease over many years, but vary annually through a seasonal component.

In this thesis I will make use of this covariance addition to describe the flight paths of birds, proposing that each flight path is composed of a repeated, habitual element, superposed with a variation specific to the individual flight. Finally each *observed* flight path will also be subject to observation noise, bringing in a third



(a)



(b)

Figure 4.6: Gaussian process regression with noiseless and noisy data. Both plots show the same measured observation values. If these observations are noiseless (plot (a)) then the posterior distribution has a low variance and is restricted to functions that pass through the observation points. If the observations include observation noise (plot (b)) then each observation is less informative and the posterior distribution is more uncertain. The posterior mean is smoother, since much of the observed variation can be explained as observation noise rather than as variation in the latent function.

covariance. The latent, unseen elements of this model will be marginalised in the same way that the process, $x(t)$ is marginalised to produce equation (46). This will produce a Gaussian distribution over the observables, the actual positions recorded on a GPS logging device.

Chapter 5

Modelling Repeated Flights

5.1 Introduction

This chapter demonstrates how Gaussian processes can be used to model the flight paths of navigating pigeons. The focus will be on data collected within an experimental paradigm where individual birds are released repeatedly from a consistent release site up to twenty times, during which time they are expected to memorise 'habitual routes' between the release site and the home loft to which they remain loyal and recapitulate with high fidelity. It will be argued that this loyalty makes a bird's future flights *predictable*. By constructing an appropriate model, predictability will be demonstrated and it will be seen how this predictability emerges over a series of releases as the bird alters from a highly variable flight pattern to a recognisable habitual route. The increase in predictability will be shown to be accompanied by greater certainty in the location of the habitual route. Finally I will discuss to what extent the habitual route is a reality as opposed to a useful element of the model and how the repeated structure of flight paths might be alternatively represented.

5.2 Data

The observed data used in this chapter and the next consists of recorded flight paths from birds released at four sites. All birds were returning to the loft at the Oxford Field Station. The data were collected for two previous studies, Meade et al.

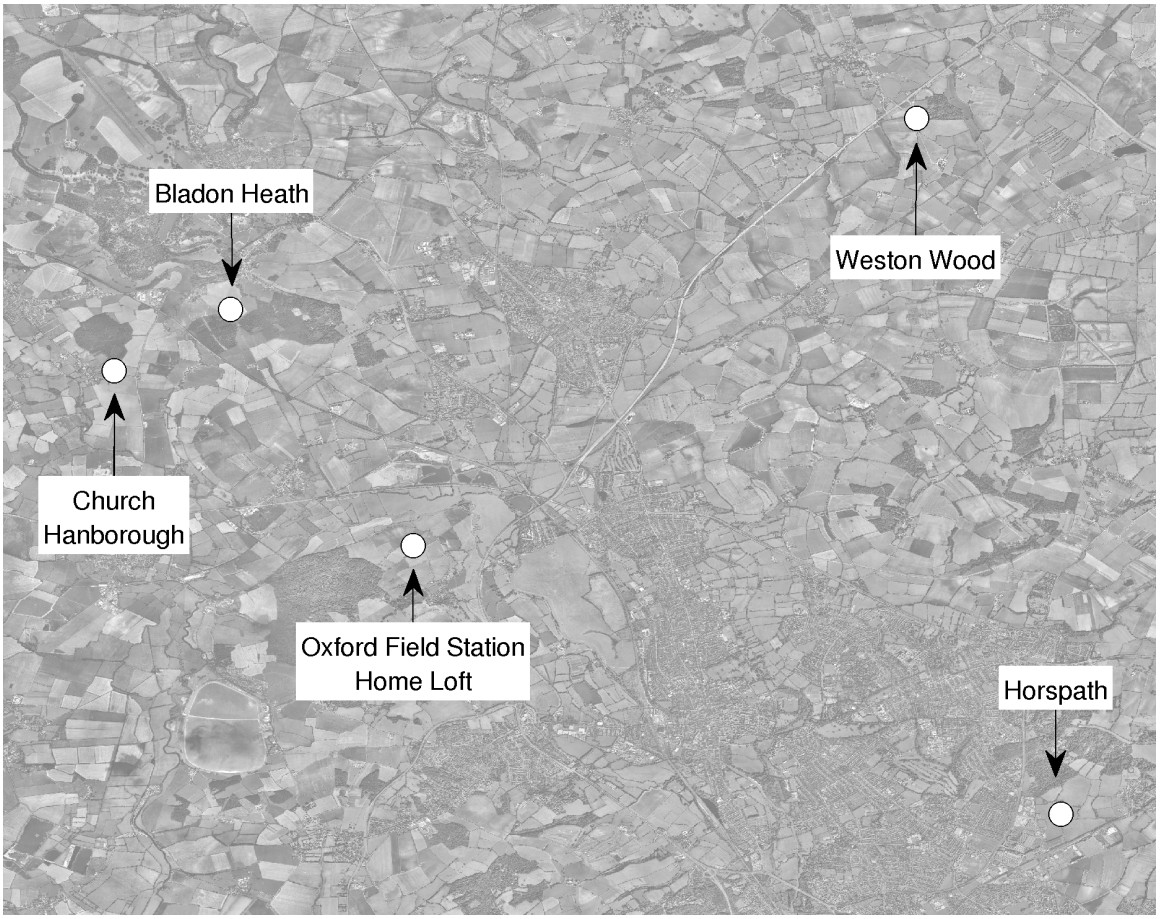


Figure 5.1: The location of the four release sites used in the repeated release experiments relative to the home loft.

[2005] and Armstrong et al. [2008]. All experiments were performed according to the protocols established by Biro [2002], Biro et al. [2002].

The four release sites were: Bladon Heath (BH, distance to home 5.0km, bearing to home 153 degrees); Church Hanborough (CH, distance to home 5.3km, bearing to home 129 degrees); Horspath (HP, distance to home 10.4km, bearing to home 300 degrees); Weston Wood (WW, distance to home 10.6km, bearing to home 221 degrees). The locations of these sites, along with the home loft, are indicated in Figure 5.1.

Across the four sites, 31 individual birds were used, 7 at Bladon Heath and 8 at each of the other three sites. Every bird completed at least 20 flights from the selected release site. Where more than twenty flights were completed only the first twenty have been included in the analysis to ensure comparable analysis across all birds.

Birds released at Horspath and Weston Wood had no previous experience of release experiments except four releases less than 1km from the loft prior to these experiments. Birds released at Bladon Heath and Church Hanborough had previously been used in release experiments from other sites, but all birds had no previous experience of the release site in this experiment prior to the first release.

Birds released at Church Hanborough were not tracked on the first flight, hence the analysis for this site begins with the second flight, but still extends only to the twentieth flight. In addition, these birds were fitted with neodymium iron boron magnets [Haugh et al., 2001]) immediately before each tracked release.

Further experimental details can be found in the original studies, Meade et al. [2005] (BH, CH) and Armstrong et al. [2008] (HP, WW).

5.3 Concept

A pigeon's loyalty to its habitual route makes it *predictable* — once route learning is accomplished we can expect subsequent flights to remain in the vicinity of the mem-

orised route. I suggest that observed flight trajectories represent imperfect attempts to replicate an *unseen and never seen* idealised habitual route. The idealised habitual route will be the model's represention of the information the bird has about its preferred route home. A fundamental assumption of this model will be that we only have access to this information via observations of true flight paths. Importantly I will make as few prior assumptions about the likely course of this route as possible, disregarding information about environment and the landscape.

I assume that variation around the idealised habitual route is non-predictable and represents correlated noise. Clearly if a bird begins its flight by heading west relative to its habitual route it will remain west of the habitual route for some duration as a result of its finite speed and acceleration. However, I assume that there is no *repeated* structure in the flight path beyond that encoded by the habitual route.

The task will be to use past observations of flight paths to predict the paths of future flights. I will specify a prior distribution of flight paths and then calculate the posterior distribution conditioned on observed flights. I will then test the quality of these predictions by evaluating the posterior for the subsequently observed test data. The quality of the model prediction will be tested by comparing the posterior log-probability of the test data with the prior log-probability of the test data. This will be a measure of how much information the training data contains about the test data.

5.4 Model

Each flight trajectory, $x_i(t)$, is a two-dimensional, continuous function of time. In this model an observed flight trajectory, $x_i(\boldsymbol{t}_i)$, represents a sample from a Gaussian process, with observations at times, $\boldsymbol{t}_i$, with a mean, $h(\boldsymbol{t}_i)$, that represents the habitual route, and a covariance, $k_\phi(\boldsymbol{t}_i, \boldsymbol{t}_i)$, that determines the scale of variation around the habitual route and the smoothness of the trajectory (parametrised by input and output scales $\phi \equiv \{\lambda, \sigma\}$). Multiple trajectories from the same bird will be assumed

to have been generated from a common idealised habitual path – mathematically this means they are identically and independently distributed from this Gaussian process, sharing a common mean function, $h(t)$, representing the idealised habitual route. The finite precision of the GPS device introduces observation error, which I model as isotropic Gaussian noise with variance η^2. Observation noise is included in the model through the addition of an identity covariance matrix, $\delta(\boldsymbol{t}_i, \boldsymbol{t}_i)$, as shown in Section 4.4.3. The resolution of a typical GPS device is within five metres, which informs the prior distribution over this hyper-parameter. With these considerations the distribution of a single flight path, $x_i(\boldsymbol{t}_i)$, conditioned on knowing the habitual route, $h(t)$, is given by,

$$p(\, x_i(\boldsymbol{t}_i) \mid h(t), \phi, \eta, I\,) = \mathcal{N}\big(x_i(\boldsymbol{t}_i); h(\boldsymbol{t}_i), k_\phi(\boldsymbol{t}_i, \boldsymbol{t}_i) + \eta^2 \delta(\boldsymbol{t}_i, \boldsymbol{t}_i)\big)\,. \tag{47}$$

Here the subscript i indexes the flight number — therefore $\boldsymbol{t}_i$ represents the vector of observation times for flight i. The input variable t is constrained to lie between zero and one, with zero representing the release and one representing collection of bird at the loft. Thus the time index of the flights is a proportion of the total flight duration.

A Gaussian process prior distribution is placed over the common habitual route, $h(t)$. I argue that having disregarded any knowledge of the environment, symmetry requires that this distribution be centred on the straight 'beeline' route, $s(t)$, between the release site and the loft. The habitual route has its own dynamical structure parametrised by the covariance kernel, $k_\theta(t, t')$, where t and t' are any time indices for the habitual route function. Similarly to the covariance for the observed flight path, the covariance kernel for the habitual route has its own hyper-parameters (input and output scales), $\theta = \{\lambda_h, \sigma_h\}$. The habitual route is an unobserved process and thus includes no observation noise, but there is still uncertainty in the value of $h(t)$ since it cannot be observed directly. At this stage the habitual path is a continuous function. The distribution is therefore a GP rather than a multi-variate Gaussian. The probability of a given habitual route is given by,

$$p(\, h(t) \mid s(t), \theta, I\,) = \mathcal{G}P(h(t); s(t), k_\theta(t, t'))\,. \tag{48}$$

Since the habitual path is never directly observed I marginalise by integration over all possible values, using Equations (A5) and (A7), to obtain a distribution over sets of flights that share a common, unknown $h(t)$. I drop the explicit dependence on the hyper-parameters θ, ϕ and η for simplicity. After marginalisation the probability of a set of paths is,

$$p\left(\begin{bmatrix} x_1(\boldsymbol{t}_1) \\ \vdots \\ x_n(\boldsymbol{t}_n) \end{bmatrix} \middle| \, s(t), \boldsymbol{\Sigma}, I \right) = \mathcal{N}\left(\begin{bmatrix} x_1(\boldsymbol{t}_1) \\ \vdots \\ x_n(\boldsymbol{t}_n) \end{bmatrix}; \begin{bmatrix} s(\boldsymbol{t}_1) \\ \vdots \\ s(\boldsymbol{t}_n) \end{bmatrix}, \boldsymbol{\Sigma}\right), \tag{49}$$

in which $\boldsymbol{\Sigma}$ is a combined covariance matrix of the form:

$$\begin{aligned} \boldsymbol{\Sigma} = & \begin{bmatrix} k_\phi(\boldsymbol{t}_1, \boldsymbol{t}_1) & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & k_\phi(\boldsymbol{t}_n, \boldsymbol{t}_n) \end{bmatrix} \\ + & \begin{bmatrix} k_\theta(\boldsymbol{t}_1, \boldsymbol{t}_1) & k_\theta(\boldsymbol{t}_1, \boldsymbol{t}_2) & \dots \\ k_\theta(\boldsymbol{t}_2, \boldsymbol{t}_1) & \ddots & k_\theta(\boldsymbol{t}_{n-1}, \boldsymbol{t}_n) \\ \vdots & k_\theta(\boldsymbol{t}_n, \boldsymbol{t}_{n-1}) & k_\theta(\boldsymbol{t}_n, \boldsymbol{t}_n) \end{bmatrix} \\ + \eta^2 & \begin{bmatrix} \delta(\boldsymbol{t}_1, \boldsymbol{t}_1) & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \delta(\boldsymbol{t}_n, \boldsymbol{t}_n) \end{bmatrix}. \end{aligned} \tag{50}$$

The three terms in equation (50) correspond to three distinct facets of the model. The first is the covariance due to the variation of each path around the habitual route. This is a diagonal block matrix since this variation is uncorrelated across different paths. The second term corresponds to the covariance associated with the shared habitual route. Finally the third term is due to the observation noise associated with measuring the pigeon's position with the GPS device. By combining these three covariance functions as discussed in Section 4.4.3 the resultant covariance structure now models all aspects of the variation in observed trajectories.

The distribution over the habitual route can now be obtained by application of

Bayes' Rule,

$$p\left(\left.h(t)\right|\begin{bmatrix}x_1(\boldsymbol{t}_1)\\ \vdots \\ x_n(\boldsymbol{t}_n)\end{bmatrix}, I\right) = \mathcal{G}P(h(t); m_h(t), \Sigma_h(t,t')) \tag{51}$$

$$m_h(t) = s(t) + k_\theta\left(t, \begin{bmatrix}\boldsymbol{t}_1\\ \vdots \\ \boldsymbol{t}_n\end{bmatrix}\right)\boldsymbol{\Sigma}^{-1}\begin{bmatrix}x_1(\boldsymbol{t}_1)\\ \vdots \\ x_n(\boldsymbol{t}_n)\end{bmatrix} \tag{52}$$

$$\Sigma_h(t,t') = k_\theta(t,t') - k_\theta\left(t, \begin{bmatrix}\boldsymbol{t}_1\\ \vdots \\ \boldsymbol{t}_n\end{bmatrix}\right)\boldsymbol{\Sigma}^{-1}k_\theta\left(\begin{bmatrix}\boldsymbol{t}_1\\ \vdots \\ \boldsymbol{t}_n\end{bmatrix}, t'\right) \tag{53}$$

Marginalisation of the unknown $h(t)$ leads to the posterior distribution of subsequent flights conditional on those already observed. Future flights may be sampled at any time indices and therefore the distribution is a GP. The probability of a new observed flight path, $x_*(\boldsymbol{t}_*)$, conditioned on previously observed flight paths is given by,

$$p\left(\left.x_*(\boldsymbol{t}_*)\right|\begin{bmatrix}x_1(\boldsymbol{t}_1)\\ \vdots \\ x_n(\boldsymbol{t}_n)\end{bmatrix}, I\right) = \mathcal{N}\big(x_*(\boldsymbol{t}_*); m_h(\boldsymbol{t}_*), \Sigma_h(\boldsymbol{t}_*,\boldsymbol{t}_*) + k_\phi(\boldsymbol{t}_*,\boldsymbol{t}_*) + \eta^2\delta(\boldsymbol{t}_*,\boldsymbol{t}_*)\big) \tag{54}$$

Bayesian marginalisation of the hyper-parameters, ϕ, θ and η is used to accurately incorporate the uncertainty associated with these variables, using Monte Carlo techniques to estimate integrals numerically (see Implementation). The prior probabilities for the hyper-parameters are assigned independent Gaussian distributions (over the log values of the hyper-parameters because their true value is bound at zero).

5.5 Implementation

The Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970] (See Appendix B and MacKay [2003] Chapter 29) was used to generate one thousand samples

from the posterior distribution of the hyper-parameters based on the training data (past flight paths). These hyper-parameter samples were then used to numerically integrate over the hyper-parameters in calculating the conditional probability of the test data (future flight paths) from equation (54) and the posterior distribution of the habitual route from equation (51). The prior probability of the test data was calculated by drawing one thousand samples from the prior distribution of the hyper-parameters and numerically integrating over the hyper-parameters in equation (49).

The covariance function used was of the Matérn class with differentiability parameter $\nu = 3/2$. This was selected by evaluating the marginal likelihood of the test data, calculated using the final 5 paths from every bird at each site. Three covariance functions were tested; the squared-exponential function and Matérn functions with $\nu = 3/2$ and $\nu = 5/2$. For 23 of the 31 birds tested the Matérn function with $\nu = 3/2$ had the greatest marginal likelihood. In the remaining cases the Matérn with $\nu = 5/2$ was favoured. The squared-exponential was convincingly rejected in all cases. It is worth noting that the results are not qualitatively sensitive to the choice of covariance function or to changes in the prior distribution of the hyper-parameters within the options tested.

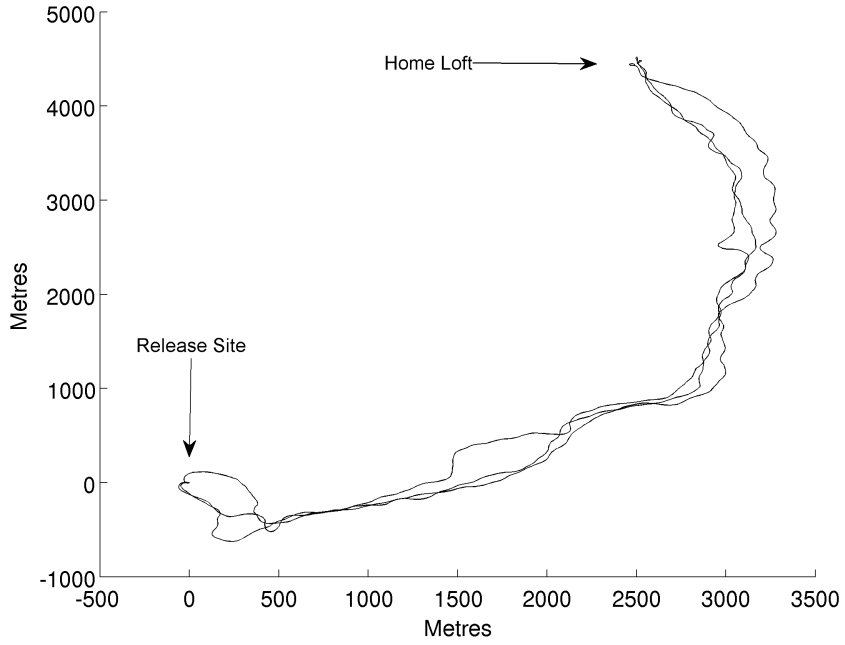
The covariance functions were conditioned on the known position of the bird at the start and end of the flight (the release point and the home loft). Paths were trimmed to remove all positions within one hundred metres of the loft, to avoid problems associated with differing release and collection points and to minimise the impact of 'circling' behaviour after the release that disrupts the temporal alignment of different paths (see *The problem with time*, section 8.1.1). To account for this trimming the conditioning on start and end points considered these to be known to within one hundred metres, using isotropic Gaussian noise to represent the uncertainty.

All time indices were normalised to begin at zero at release and end at one at the loft. When trimmed paths were used the first untrimmed point was labelled as zero and the final untrimmed point was labelled as one. Thus time indices represent the proportion of the total flight duration.

Figure 5.2 shows an example of the hyper-parameter sampling. The model was trained on three paths from a single experienced individual, shown in panel (a). Panel (b) shows samples from the prior distribution of the hyper-parameters as input-scale, output-scale pairs. Samples are shown for the hyper-parameters associated with variation of the observed paths around the habitual route (Φ) and for the variation of the habitual route around the straight line path (Θ). Panel (c) shows samples for the same hyper-parameters from the posterior distribution, sampled using the Metropolis-Hastings algorithm. It is clear there is a very substantial reduction in the posterior uncertainty associated with these parameters, which are well defined by the data and hence not strongly sensitive to changes in the prior distribution. Samples were taken after a 'burn-in' period of 100 samples, by which time inspection of the sample chains from various initial positions showed convergence to the same sampling distribution (see Appendix B).

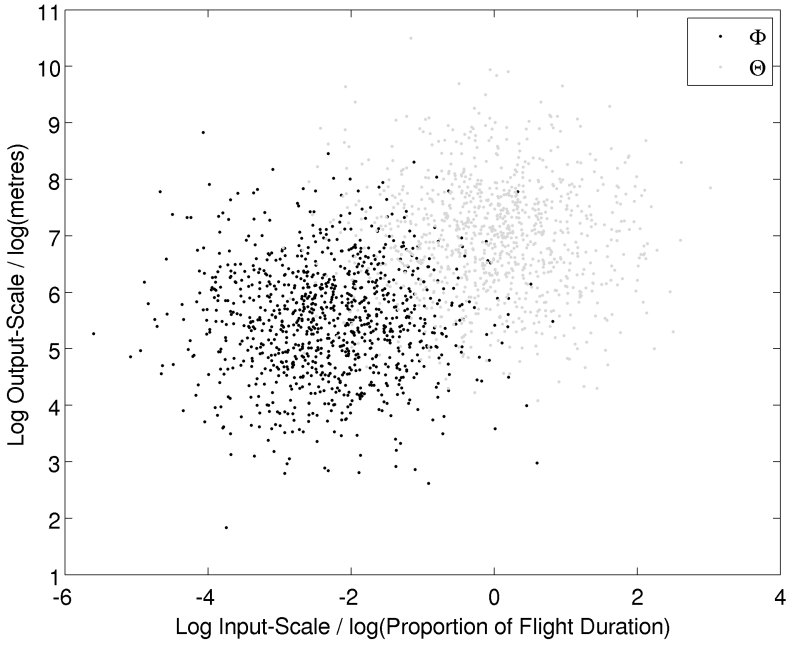
5.6 Results

Figure 5.3 gives a graphical representation of the naive prediction before observing any training data, described by equation (49), at the Bladon Heath release site. This represents the predicted 'flight corridor' prior to observing any flight paths. The distribution is symmetrical about the straight line between release point and the home loft (release point and loft indicated by white markers), while the black dashed lines indicate the first standard deviation. The width of the distribution is composed of two elements. The first element is our uncertainty over the habitual route, indicated by the dashed red lines. Before observing any flight paths this is very large since there is no indication of the likely route except that it will start at the release point and finish at the home loft. This uncertainty is reduced after observing flight paths. The more consistent those flight paths are the lower the uncertainty over the habitual route. The second element is the uncertainty due to the variation of observed flight paths around the habitual route. This is expected to be small *a priori* since the model



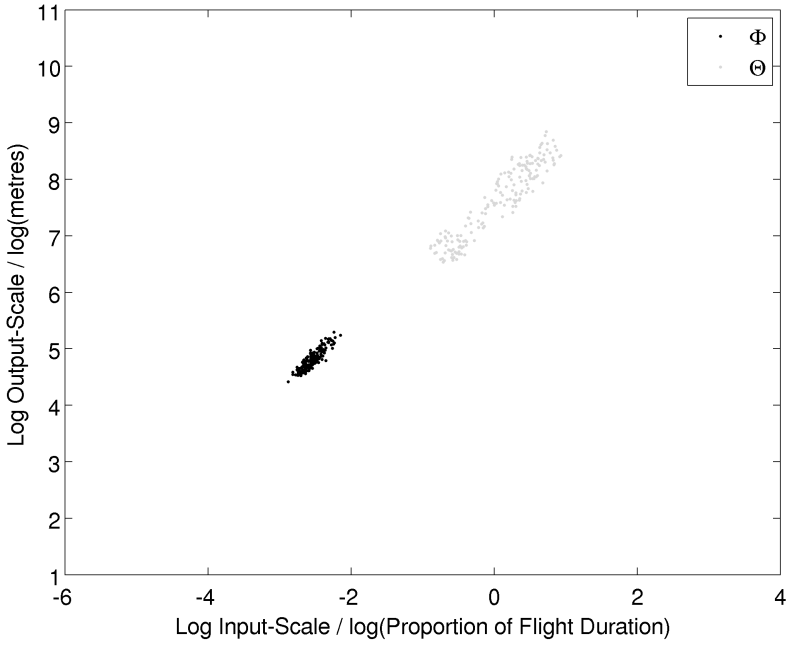
(a)

Figure 5.2: Prior and posterior distribution of hyper-parameter samples. Panel (a) shows example test data used to infer the posterior distribution of the hyper-parameters. Panel (b) shows hyper-parameter samples generated from the prior distribution, with each logarithmic input-scale plotted against the corresponding logarithmic output-scale. Panel (c) shows samples generated from the posterior distribution using the Metropolis-Hastings algorithm and the test data in panel (a).



(b)

Figure 5.2: ...Continued



(c)

Figure 5.2: ...Continued

is based on the bird's loyalty to the habitual route. As flight paths are observed the model can estimate the amount of variation around the habitual path, which will be lower the more consistent the flight paths are.

Figure 5.4 indicates the model trained to two different birds using Equation (54), based on a set of five training paths from each bird. The uncertainty associated with the future flight paths is dramatically reduced, as evidenced by the lower standard deviation, while the mean predictions, indicated by the heavy black lines, now follow the large scale pattern of the training paths. The width of the two distributions accurately reflect the scale of intrinsic variation in the birds' flight paths. After five observed flights the uncertainty in the habitual path is low and the width of the predictive distribution is dominated by the intrinsic variation around the habitual path. I took consecutive pairs of flight paths and, using equation 54, predicted the path of the next flight (e.g., predicting the path of the third release based on the paths of the first two flights). I compared this with the prior probability of the subsequent trajectory to give a metric of predictability using Marginal Information Gain (MIG), defined as:

$$\text{MIG/bits} = \log_2 p(x_i(\boldsymbol{t}_i) \,|\, x_{i-1}(\boldsymbol{t}_{i-1}), x_{i-2}(\boldsymbol{t}_{i-2}), M) - \log_2 p(x_i(\boldsymbol{t}_i) \,|\, M), \qquad (55)$$

in which M represents the model described above, $x_i(\boldsymbol{t}_i)$ is the flight path being predicted and $x_{i-1}(\boldsymbol{t}_{i-1})$ and $x_{i-2}(\boldsymbol{t}_{i-2})$ are the two most recent flight paths before the predicted flight path. Values of MIG above zero indicate predictable behaviour — the flight is more likely in the light of observations than it was *a priori*. I use two flights to make predictions since the model is required to learn both the scale of variation around the straight line prior and the scale of inter-flight variation. Figure 5.5 shows the MIG averaged over the 31 birds as a function of the flight number being predicted. The clear trend is for increasing predictability which corresponds to our belief that the bird is initially naive and forms a memorised route. Figure 5.6 shows that this is accompanied by a reduction in the corresponding uncertainty on the habitual route inferred from successive pairs of flights as defined by equation 51. The clear increase

Figure 5.3: The naive prediction for flight paths before model training. The heavy black line represents the mean of the distribution, with one standard deviation indicated by the dashed black lines. The dashed red lines indicate one standard deviation on the distribution of the habitual path, which shares the same mean.

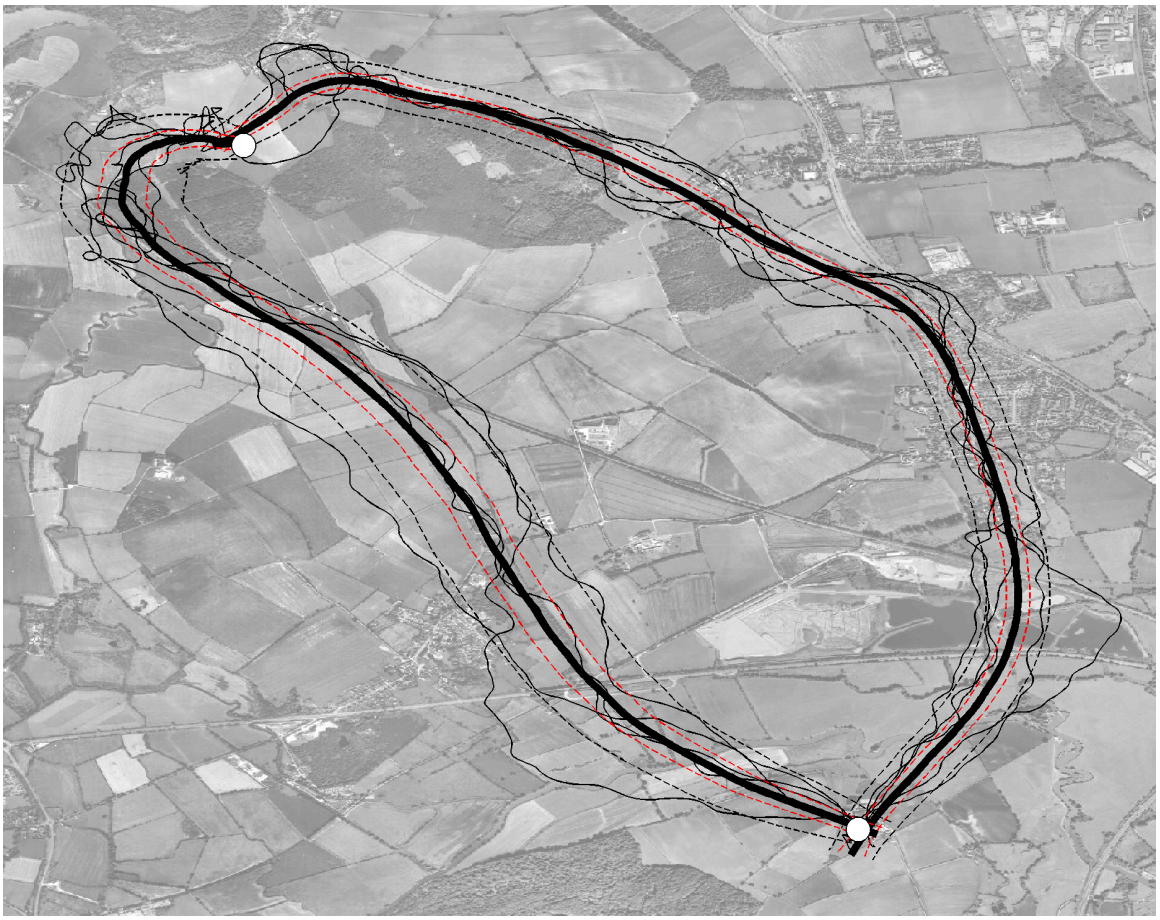


Figure 5.4: Model predictions when trained separately on two sets of training paths (light black lines). The two distributions are represented by the mean, standard deviation and the standard deviation on the mean as in Figure 5.3

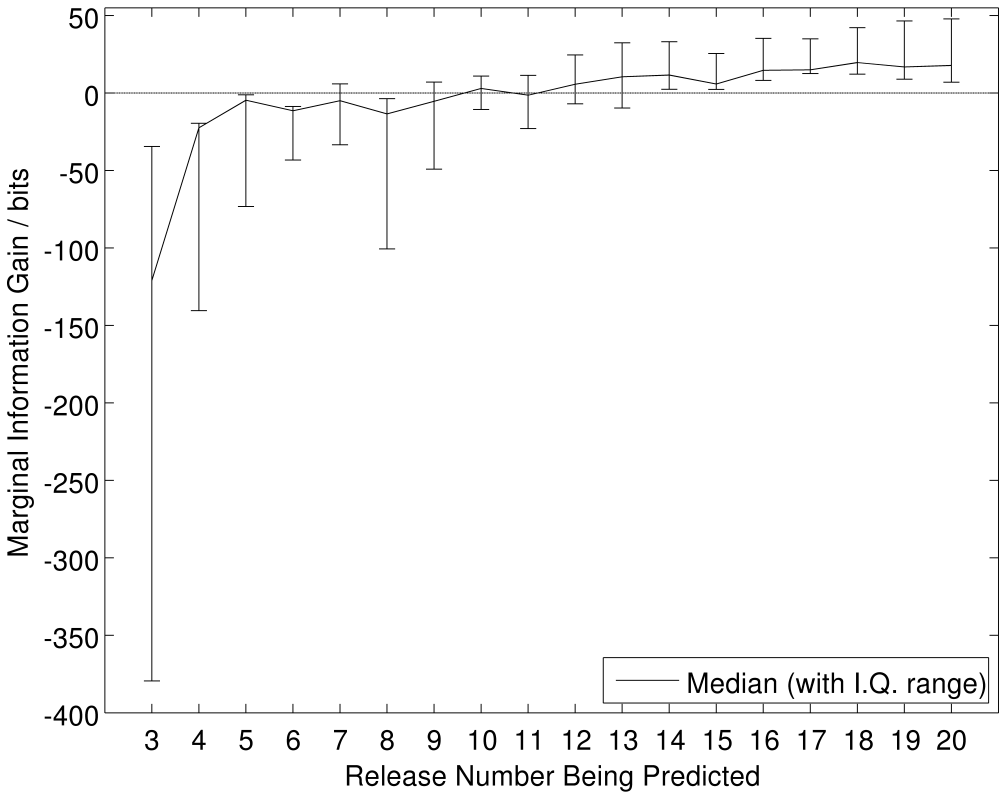


Figure 5.5: Marginal Information Gain in predicting each flight path from its two immediate predecessors. The indicated values represent the median value from 31 birds. The error bars represent the inter-quartile range.

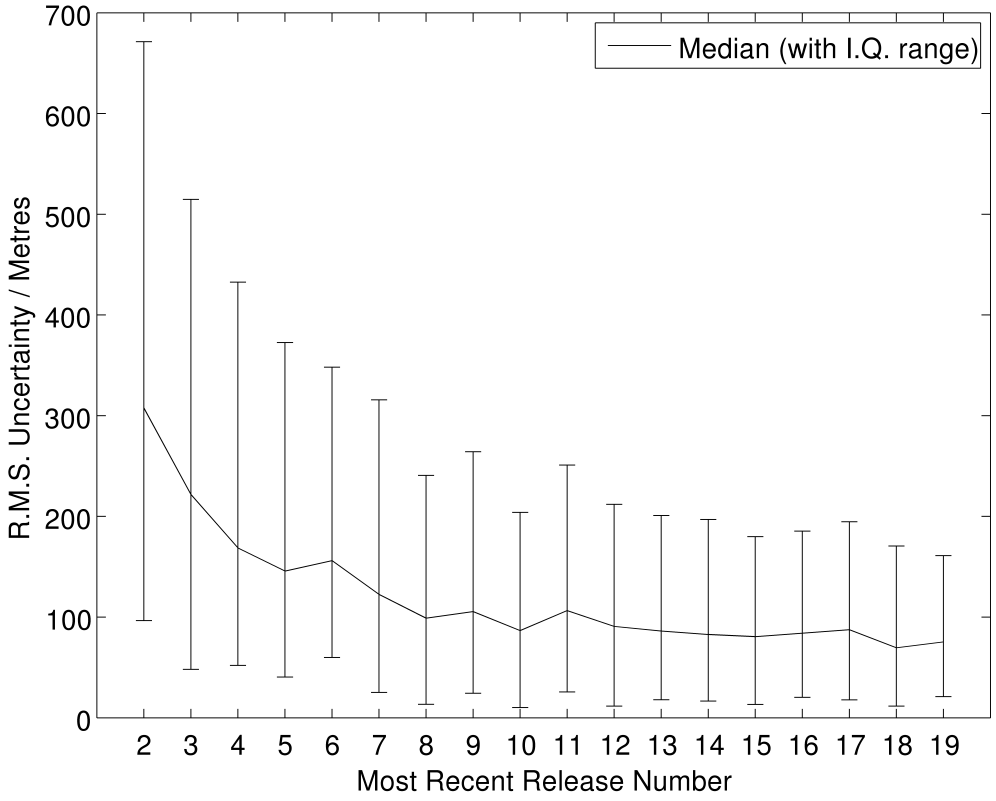


Figure 5.6: Average (RMS) spatial uncertainty for the location of the idealised habitual route, as inferred from consecutive pairs of flights. The indicated values represent the median value from 31 birds. The error bars represent the inter-quartile range

in predictability in Figure 5.5 is a confirmation of route-learning behaviour from a new perspective. The very low predictability of the first few flights is a consequence of the extremely variable nature of the first flight of many birds. Earlier studies have shown that birds released at unfamiliar sites tend to circle the release site after being released and subsequently home along high disordered trajectories (see figures in [Meade et al., 2005]). After only one previous experience of the release site this effect is often greatly reduced, explaining the substantial increase in predictability in the first few flights. After this early naivety is overcome the trend becomes a steady but gradual increase in predictability continuing until the final flight.

The development of the habitual route can be seen directly in Figure 5.6, which shows the R.M.S. standard deviation in the position of the habitual route, as inferred from Equation (51) using the same pairs of paths used to make predictions of the subsequent path. Because every inference of the habitual route is made using only two flights the increasing precision in the inference is likely due to the greater route-knowledge of the birds rather than the learning process of the model. This asymptotically declining curve mirrors the previous measures of route-learning, which show an asymptotically declining distance between successive paths or groups of paths [Biro et al., 2004, Meade et al., 2005, Armstrong et al., 2008].

These results can be broken down into specific release sites, as shown in Figure 5.7. Increasing predictability is observed at all four of our test sites. However, the results indicate key site-specific differences. The very poor predictability of early flights is evident at all sites with the the exception of Weston Wood (plot (d)). This is likely to be a consequence of a site-specific factor, the large road leading from this release site in the direction of the home loft (see Figure 5.1). The attraction of the birds to this extremely prominent feature reduces the variable nature of early flights. Another notable feature of the results is that the low early predictability endures for longer at Bladon Heath and Church Hanborough (plots (a) and (b)). Inspection of the final flights at these sites suggest that at each two principle routes develop. In early flights birds are sometimes observed to 'trial' each of these routes. Thus the

enduring low predictability may be indicative of this exploration stage. By contrast at Horspath and Weston Wood (plots (c) and (d)) the birds have no discernible clustering in their choice of habitual route and progress more smoothly to their individual choice. After roughly 10 releases all four sites exhibit similar patterns of gradually increasing predictability and similar scales of information gain.

5.7 Discussion

Visual inspection of recorded flights suggests increasingly predictable behaviour centred on loyalty to a habitual route. In this chapter I have shown that this is reflected in increasing predictability as quantitatively defined using probability theory. There are increasingly tight bounds on the location of an 'idealised habitual route' that represents the encoded memories the bird uses for navigation. This supports the route learning hypothesis and demonstrates more clearly than the predictability result the asymptotic nature of the learning curve.

Route learning is a confirmed result in the zoology literature [Biro et al., 2002, Biro, 2002, Biro et al., 2004, Meade et al., 2005, 2006, Biro et al., 2006a]. The power of reproducing this result in the manner of a probabilistic model will become more apparent in the subsequent chapters, however it is worth noting what this analysis has already revealed that was previously unavailable. Using the new analysis it is possible, for the first time, to place a probability on an observed flight path. Rather than estimating the distribution of key metrics such as vanishing bearings it is possible to estimate (and rigorously test using Bayesian methods) the distribution of the flight paths themselves. The analysis of experimental interventions boils down to the question 'has there been a change in behaviour?' We can now assess the probability of flight paths generated *after the intervention* in the light of those generated before. This enables the flight paths to become the *fundamental unit* of observation and analysis.

So far I have built a model based on a hypothetical idealised habitual route that

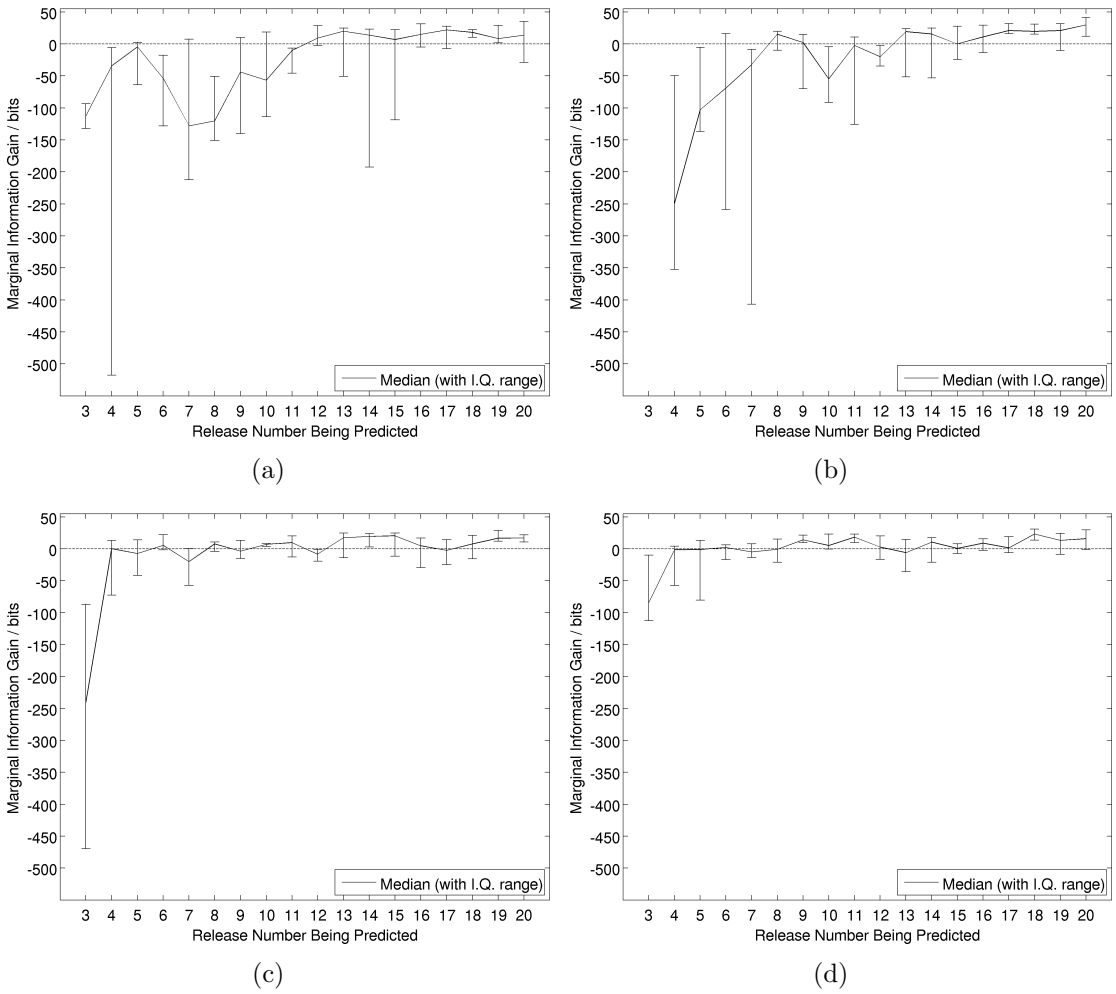


Figure 5.7: Increasing predictability of flight paths, site-by-site, showing marginal information gain as a function of the release number of the predicted flight. Sites are (a) Bladon Heath (7 birds); (b) Church Hanborough (8 birds); (c) Horspath (8 birds); (d) Weston Wood (8 birds). Figures show the median values and the inter-quartile range.

represents the route the pigeon attempts to mimic. In practice however, it is unlikely that the birds encode their navigational memories in the form of a continuous smooth geo-stationary route. To do so would be an enormous waste since a very similar route could easily be reproduced with far fewer, discrete memorised locations. There would surely be little efficacy in memorising substantially more information than is required to reproduce the route within the observed inter-flight variability. An alternative hypothesis is that birds memorise a set of *waypoints* that serve to encode the necessary information to fly home. These would be discrete, geo-stationary locations that the bird visits successively on each flight. It is clear that from such behaviour would emerge the resemblance to a continuous memorised route since the deviation between each waypoint would be small if they were sufficiently close, and conversely waypoints would need to be sufficiently close to prevent excessive variation as this might lead to a failure to locate the next waypoint.

The next chapter will further develop the model described in this chapter. The habitual route will be replaced by discrete waypoints. The power of the method will become apparent as predictability becomes the metric whereby potential sets of waypoints are judged to create an automated waypoint detection algorithm. Detecting and using waypoints will allow us to probe closer to the true representation of the birds 'familiar area map'.

Chapter 6

Landmark and Waypoint Identification

6.1 Introduction

The high-fidelity loyalty of birds to their idiosyncratic habitual route has been offered as support for the hypothesis of a primarily visual navigation system in the familiar area [Biro et al., 2004, Meade et al., 2005]. In the last chapter I explored the concept of the habitual route as a means of predicting future flight paths. I concluded by arguing that the habitual route, while a useful concept, is potentially the result of a simpler and more fundamental basis of navigational memories based on waypoints associated with visual landmarks. Under this hypothesis the bird *pilots* from one waypoint to the next until the target is reached.

In this chapter I apply the same GP model as before, but explore the potentially discrete nature of the birds' memories by assessing the predictive power of subsets of the data.

6.2 What is a waypoint?

Chapter 2 discussed the different types of potential navigation strategies a bird might use to reach home, and Section 2.7 showed there was strong support for navigation based on *pilotage* between visual landmarks within the familiar area. Using such a strategy, the bird flies from one landmark to the next in order, along a chain of

landmarks that eventually leads to the final destination. These landmarks act as *waypoints*, geo-stationary locations that the bird physically visits, or flies close to, during flight.

It is important to distinguish technically between a landmark and a waypoint. While a waypoint is likely to be associated with a landmark it is by no means guaranteed that the bird identifies the location in a way similar to a human. Conversely it is possible that a landmark may not constitute a waypoint. I will use the term waypoint to mean a location the bird physically visits. Therefore a bird might use a distant object as a landmark to determine its course but that would not be associated with a waypoint.

As well as consisting of a recognisable geo-stationary location a waypoint may have other properties. Principally it may be associated with a bearing indicating the direction to the next waypoint. It is unknown whether birds tend to use waypoints that are visible from each other (at flight height) or whether they learn a bearing to the subsequent, initially unseen, waypoints. Such a bearing could be identified by the arrangement of nearby elements of the landscape or by using one or more of the 'compass' mechanisms discussed in section 2.4. More speculatively a waypoint may have a (possibly vague) temporal element - the bird may have some idea how long it should take to reach the next waypoint and use this to determine if some effect has taken it off course.

6.3 How can waypoints be detected?

The first step in waypoint identification is the definition of a waypoint as a location the bird physically visits. Candidate waypoints can be identified by visual inspection as locations that repeated flights visit with high frequency and low variation. Visits to waypoints may be associated with low variability between different flight paths since at these times the bird should be most certain about its position. A subjective assessment can hence be made by searching for this high fidelity return to previously

visited locations. Among these candidates, we will also be more confident about those that appear unlikely to be the result of chance. A few repeated visits to a location directly between the release site and the home loft may be suggestive, but will be far less surprising than a more consistent loyalty to a location far from the most efficient possible route. A potential method for objectively determining whether a putative waypoint should be accepted would be to infer how unlikely the bird was to have visited a location, under the null hypothesis that no waypoint existed. Sufficiently unlikely clustering of paths in small regions of space, far from the expected route, would be accepted as evidence for the rejection of this hypothesis.

A significant obstacle to an algorithmic detection of waypoints lies in the lack of training data available. A standard machine-learning approach to classifying points of interest would be to learn the 'signature' effects of those features from a set of 'labelled' data. In this case that would consist of flight paths using known waypoints, the locations of which would be marked in advance. An detection algorithm could then learn what characteristics of the flight path were strongly correlated with these locations. By looking for this signature in other flight paths it would estimate the locations of the waypoints used in these. Unfortunately, although it is possible to hypothesise what effects waypoints will have (e.g., lower variation, sharp decreases and subsequent increases in variation, potentially stronger turning angles, high tortuosity or entropy, etc.) this can not be confirmed from a labelled data set.

Here I propose a different approach. It is not necessary to guess at the effects waypoints will produce on the observable flight paths. Nor need the analysis be confined to isolated sections of those flight paths — if waypoints encode the bird's full knowledge of its route they ought to have effects throughout the flight. Instead of learning a characteristic pattern or signature associated with a waypoint it is better to return to first principles and construct a predictive model for the flight paths that includes waypoints.

This approach changes the problem of identifying waypoints from a *supervised learning* problem to an *unsupervised learning* problem; instead of learning to identify

waypoints from previously observed examples, waypoints are posited as a fundamental feature of the data, which will be revealed by looking for the most efficient encoding of the data. In essence, waypoints will be regarded as a low dimensional encoding of very high (infinite in the case of continuous functions) dimensional flight path data.

6.4 Optimal predictors of flight paths

Instead of looking for a signature effect on the flight path such as low variation, high entropy or beaconing, I approach the problem of identifying waypoints in reverse. Rather than using the flight paths to identify waypoints I use putative waypoints to predict the flight path. This then poses the waypoint identification problem as a classic inverse problem, requiring one to determine an optimum set of waypoints that maximises the probability of the observed data.

I make the assumption that a bird's knowledge of its route home is encoded *solely* by the location of its waypoints. This is the basis of the pilotage hypothesis. Waypoints could also potentially include a bearing to the next waypoint, though I have so far excluded this possible extension. Any set of putative waypoints can be described as a list of geographical locations, probably associated with an ordering (visit waypoint A, *then* visit waypoint B). Depending on how consistent the bird's flight is, each waypoint may potentially have a weak temporal association, i.e., the bird expects to reach a waypoint after a certain time. I will make the further assumption that waypoints are reached at the same time on each flight as a proportion of the entire flight duration. This is a reasonable approximation since I will be looking at flight paths generated after a large number of training flights once the habitual route has been established and is stable. Moreover, variation in flight speed within any single flight is relatively low (roughly 20-25% of the average flight speed). Nonetheless it is clear that this *is* an approximation and its implications and potential solutions will be discussed at the end of this chapter.

Given the assumptions outlined above I can construct a method for predicting

flight paths from a series of landmarks. Equation (54) shows how to predict a new flight path given previously observed flights. In that equation the time indices of the previously observed flights were taken to represent all of the times at which there were recorded positions for the bird via the GPS logging device. However, the equation applies equally to any arbitrary set of time indices for which there are recorded positions. I am free therefore to make predictions based on any subset of the observed data that I choose using the same equation. Since I have assumed that each waypoint is linked to a time index any set of time indices corresponds to a set of potential waypoints. The idea behind the optimal predictor method is to determine the set of time indices that make the best predictions according to Equation (54). The logic behind that aim is that those time indices correspond to the bird's waypoints since the bird is presumed to reconstruct its route home from the waypoints and the selected time indices and associated positions allow us to make the best reconstruction of that route.

To determine the location of waypoints I create a metric of predictability similar to the Marginal Information Gain metric in Equation (55). Taking a set of n paths, $\{x_1(\boldsymbol{t}_1), \ldots, x_n(\boldsymbol{t}_n)\}$, I predict each path using a subset of the other paths. Let $x_i(\boldsymbol{t}^m)$ refer to m observations of flight i at times $\boldsymbol{t}^m$. At each iteration another observation is added to the subset to maximise the MIG, defined as:

$$\mathrm{MIG}(\boldsymbol{t}^m) = \sum_i \log_2 p(\, x_i(\boldsymbol{t}_i) \mid \{x_{\bar{i}}(\boldsymbol{t}^m)\}, M\,) - \log_2 p(\, x_i(\boldsymbol{t}_i) \mid M\,), \tag{56}$$

where $\bar{i}$ indicates all considered paths except path i (I permute the calculation over all paths so as to ensure the minimum impact of outliers). In this calculation the distribution of the hyper-parameters, ϕ, θ and η, is not inferred from the subset of data, since this would force the algorithm to choose waypoint locations so as to minimise the uncertainty on the hyper-parameters, rather than capturing the spatial information that waypoints should provide. For example, the selection algorithm may select a series of closely spaced data points in order to minimise uncertainty over the input-scale hyper-parameter associated with the variation of the flight path around

the habitual route, which typically varies over short time scales. This would provide little spatial information. I aim to mimic as closely as possible the way the pigeons use a set of geographical locations to reproduce their habitual route. A pigeon does not need to use these positions to 'learn' model hyper-parameters, since these represent flight characteristics intrinsic to the bird (e.g., that it cannot move without error from one waypoint to the next in a straight line). In contrast the spatial structure of the waypoints represents real information learnt by the pigeon and encoded in its memory.

Therefore I marginalise over the hyper-parameters using the prior distribution. Note that similar results are obtained by marginalising over the hyper-parameters using the posterior distribution inferred from the complete data set, suggesting that the algorithm is not overly sensitive to the hyper-parameter distribution.

Each element of the sum in Equation (56) is the log-probability of a flight path, given a set of parameters (the waypoint locations). The MIG metric therefore acts as the log-likelihood of a given set of waypoint locations, with a constant factor given by the prior log-probability of the flight paths which does not change as a function of the waypoint locations. This allows for Bayesian inference over the choice of waypoints.

The optimum model order can be inferred by using the Bayesian model order selection techniques described in Chapter 3. It is not possible to marginalise over all sets of m waypoints for m greater than three or four, simply because the combinatorial explosion in the number of different possible sets makes it impossible to adequately explore the parameter space. However, the Forward-Selection algorithm used to select the waypoints (see Section 6.5) does not choose from the space of all sets of size m, but instead iteratively chooses one more waypoint, assuming that all those already selected are now known and fixed. Therefore, at each iteration, when choosing the next waypoint, I simply marginalise over the time index of the next waypoint and compare the marginal likelihood of this extended set of waypoints to the likelihood of the set of waypoints already selected in previous iterations. In doing this I am comparing two models — M_0: The set of waypoints already selected; M_1: The set of

waypoints already selected and one additional 'free' waypoint. The Bayes factor for M_1 over M_0 is therefore given as,

$$\begin{aligned} \mathrm{BF} &\equiv \frac{P(\,\{x_1(\boldsymbol{t}_1),\ldots,x_n(\boldsymbol{t}_n)\}\mid M_1\,)}{P(\,\{x_1(\boldsymbol{t}_1),\ldots,x_n(\boldsymbol{t}_n)\}\mid M_0\,)} \\ &= \frac{1}{N}\sum_{t_{new}=1}^{N}\exp\big(\ln 2\times[\mathrm{MIG}(\boldsymbol{t}^m,t_{\mathrm{new}})-\mathrm{MIG}(\boldsymbol{t}^m)]\big), \end{aligned} \tag{57}$$

where N is the number of possible landmark locations (restricted this to 100 in this implementation) and $\{x_1(\boldsymbol{t}_1),\ldots,x_n(\boldsymbol{t}_n)\}$ is the complete data set. If BF is less than 1 the evidence favours M_0, showing that the addition of a new waypoint would make the model predictions less accurate. Therefore I do not add the next waypoint and the algorithm stops.

The information gain metric of equation (56) is used in preference to alternative approaches to selecting data subsets in the GP literature, such as those proposed by Lawrence et al. [2003] and Seeger [2003]. These alternative metrics emphasise selecting subsets that minimise the variance of the predictive distribution, assuming a stationary GP. While the underlying GP for the model is stationary, selection of data subsets can introduce non-stationarity by reducing variance in areas around the selected data while leaving variance high in other regions. For the practical problem of modelling navigating birds this is preferable, since higher variance is expected in regions furthest from any waypoints. Since there are repeated flights I have the luxury of using observed data to judge the predictions of this model, rather than using just the predictive variance. The MIG metric also acts as a likelihood, enabling a Bayesian approach to inferring the model order (the number of waypoints) through marginalisation.

6.5 Implementation

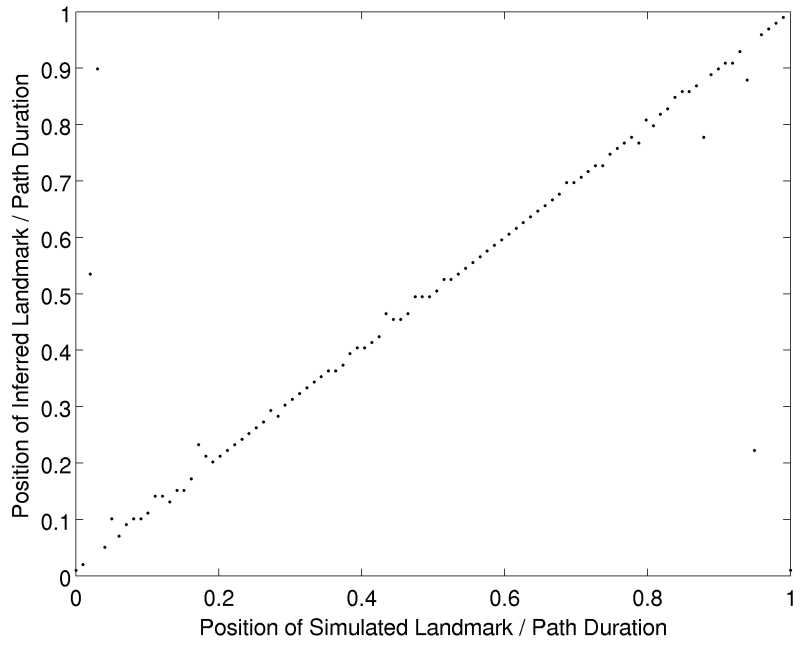
The implementation followed the procedures adopted in Chapter 4. In addition, each path was initially normalised to a preset number of points at a fixed set of proportional time indices. The raw paths did not necessarily have a large number of proportional

time indices in common, since they were of different durations and were all sampled at 1Hz. Therefore normalisation involved interpolating between existing data points to estimate the path position at the fixed indices. A GP regression fit was used to perform this interpolation, which behaved almost identically to a spline interpolation used for comparison. Paths were normalised to one hundred proportional time indices which was lower than the number of raw data. The high density of the raw data meant that uncertainties in the regression to the new time indices could be ignored for practical purposes.

6.6 Testing

No labelled data exists at present for real flights indicating the presence of known waypoints. Therefore testing of the algorithm was performed on artificially generated data. A one-waypoint set of data was produced, with the time index of the simulated landmark being increased between zero and one. At each iteration five paths were generated, using the posterior parameters indicated in Section 5.5. The paths were generated independently, except that they were constrained to pass through the straight line path at a single, common time index, thus simulating a waypoint on the straight line path. The algorithm was then run to find the first identified waypoint, which was recorded. Figure 6.1 (a) shows an example of simulated data, alongside a plot showing the first identified waypoint index as a function of the simulated waypoint index (b). The correct waypoint is identified in almost every trial. The few instances where the waypoint is incorrectly identified occur when the simulated waypoint is close to the release site or the home loft. The paths are already constrained in these regions and therefore the waypoint has less discernible influence on the paths. Placing the waypoint on the straight line between the release site and the home loft is not only a simple choice but also makes the test stronger, since waypoints here are the hardest to identify, as the path has the highest probability of being in this region by chance alone.

(a)



(b)

Figure 6.1: Testing the waypoint detection algorithm. Plot (a) shows an example of simulated data. A waypoint has been simulated by forcing the simulated paths to cross the straight line path at $t = 0.5$. Plot (b) shows the time index of the first waypoint identified by the algorithm as a function of where the simulated waypoint is placed.

6.7 Results

Figures 6.2 and 6.3 show an example of application of this algorithm over data for a single bird. Figure 6.2 shows the identified landmarks, along with the five flight trajectories used to identify them, plotted on an aerial image of the underlying landscape. I have subjectively identified features in the landscape at the landmark sites by visual inspection. Figure 6.3 shows the logarithm of the Bayes factor and the MIG metric as functions of the number of landmarks. The number of landmarks is determined by the first point at which the logarithmic Bayes factor is below zero, as this indicates that the addition of the next landmark reduces the marginal likelihood of the data. The maximum likelihood estimate for the number of landmarks occurs at the peak of the MIG function and is significantly greater than the Bayesian estimate.

Inspection of the identified landmarks points to a number of striking visual features in the vicinity. The first identified landmark is above the village of Yarnton, which is positioned at the apex of the flight trajectory and therefore does most to define the shape of the habitual route. Further landmarks are positioned over Bladon village, near the release site, and along the boundaries of the forests between Bladon and Yarnton. This corresponds to known behavioural facets of pigeon orientation. Wallraff [1994] showed that pigeons released from unfamiliar sites showed a directional bias towards villages and forests in the vicinity, and pigeons are known to avoid *crossing* forested areas because of an increased risk of predation [personal correspondence with Dora Biro and Tim Guilford]. Therefore a viable explanation for this route and the associated landmarks is that the pigeon was initially attracted towards the village at Bladon. Once there it was prevented from directly flying towards the home loft by the obstacle of the forests and therefore flew along the boundary of these. This brought it into visual range of Yarnton village, to which it was attracted before flying home. These initial biases then selected the regions of the landscape from which it could select navigational landmarks.

Figure 6.4 shows all the identified landmarks from the 31 birds used in this

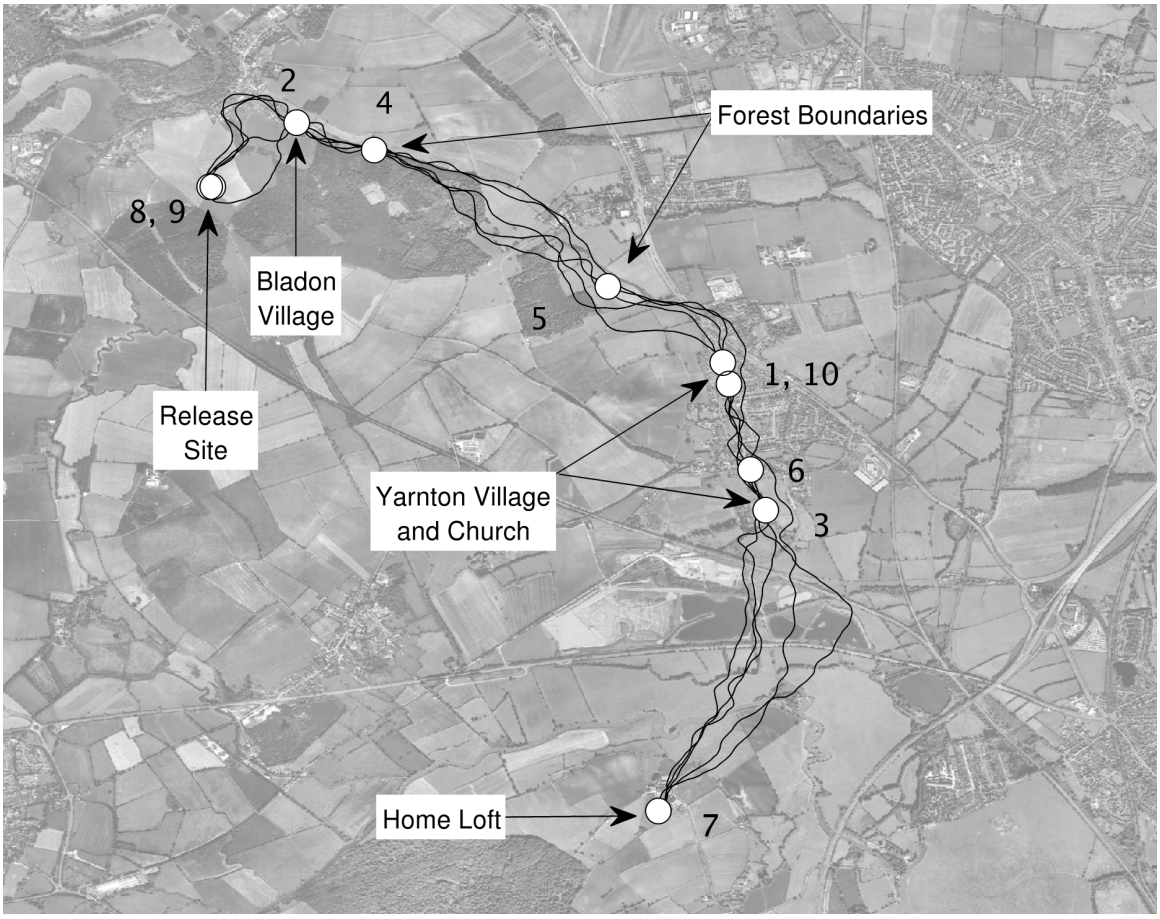


Figure 6.2: Identified waypoints (white circles) from five flight paths (black lines) from the Bladon Heath release site. Numbers from one to ten indicate the order in which the waypoints were identified, one being the first identified and ten being the last.

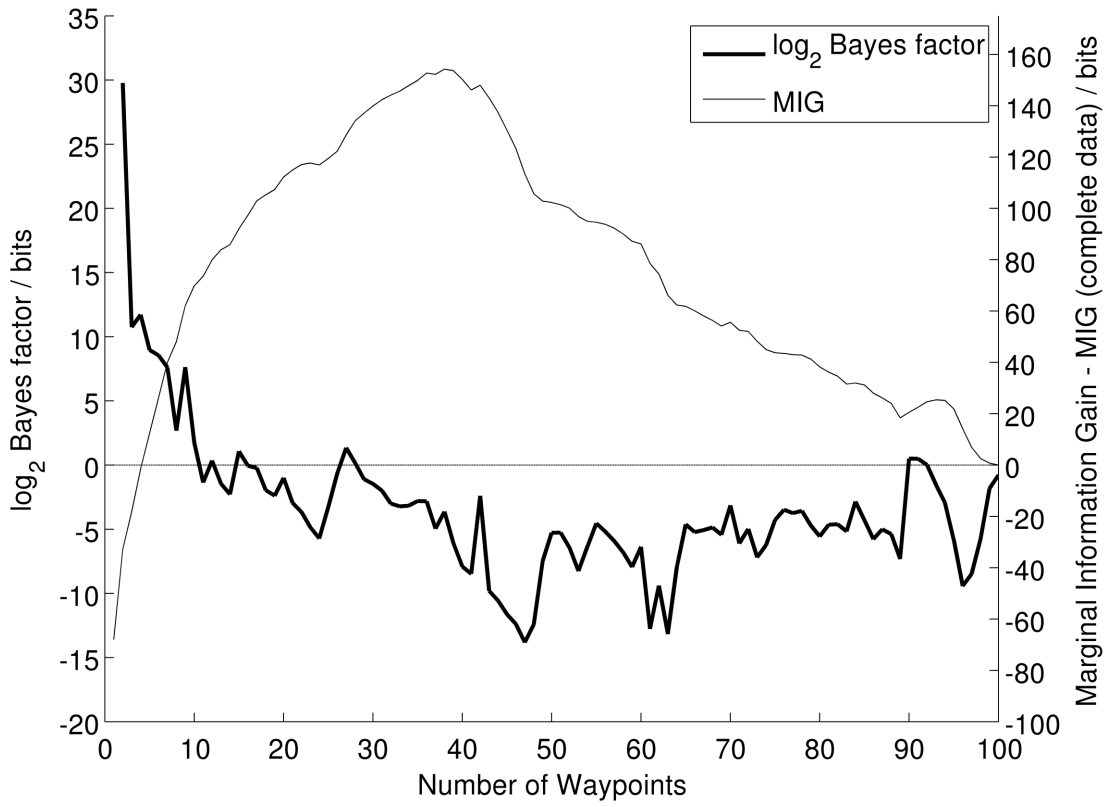


Figure 6.3: Marginal Information Gain and Bayes factor for sequentially adding waypoints in the Bladon Heath case study. The dark line represents the logarithmic Bayes factor for the addition of each waypoint. The optimal number of waypoints is selected when the log Bayes factor is first below zero. The light line represents the Marginal Information Gain in proportion to the MIG when using all data points.

study, colour coded according to the release site. The four release sites and the home loft are indicated. Figure 6.5 shows the same landmarks further broken down into each specific release site and colour-coded according to which bird they were identified from.

A notable feature of these images is the relatively low density of landmarks over urban areas. As noted in the case study, pigeons seem attracted towards small urban areas, such as villages. However, they seem either unwilling to cross them or, if avoiding them entirely is unfeasible, they form very few landmarks within them. The release site at Horspath was originally selected to explore the behaviour of pigeons over urban landscapes by forcing them to fly over the suburbs and centre of Oxford [Armstrong et al., 2008]. As can be seen here the density of landmarks is very low in suburban Oxford, near the release site, increases slightly in the centre of the city, and increases more dramatically once the trajectories leave the city and enter the rural area between Oxford and the home loft.

The identified landmarks from the Weston Wood release site reproduce earlier findings that pigeons attend to the major road leading from the release site in the direction of the home loft. The highest density of landmarks occurs when the road changes direction – at this point most of the pigeons' habitual routes leave the road and become more variable. Again there is an absence of landmarks within the urban area that intersects the natural flight corridor.

The landmarks close to the release site at Bladon Heath are a notable example of the effect of predation risk. These pigeons form landmarks all around the edge of the forested area close to the release site, and through a narrow unforested partition, but rarely fly directly across the forested areas.

The landmarks identified from releases at Bladon Heath and Church Hanborough show a substantial element of overlap, as do landmarks identified from the other sites within the region around the home loft where the routes converge. This is persuasive evidence that some underlying feature of the landscape is sufficiently visually arresting to attract not only different pigeons, but also pigeons released from different

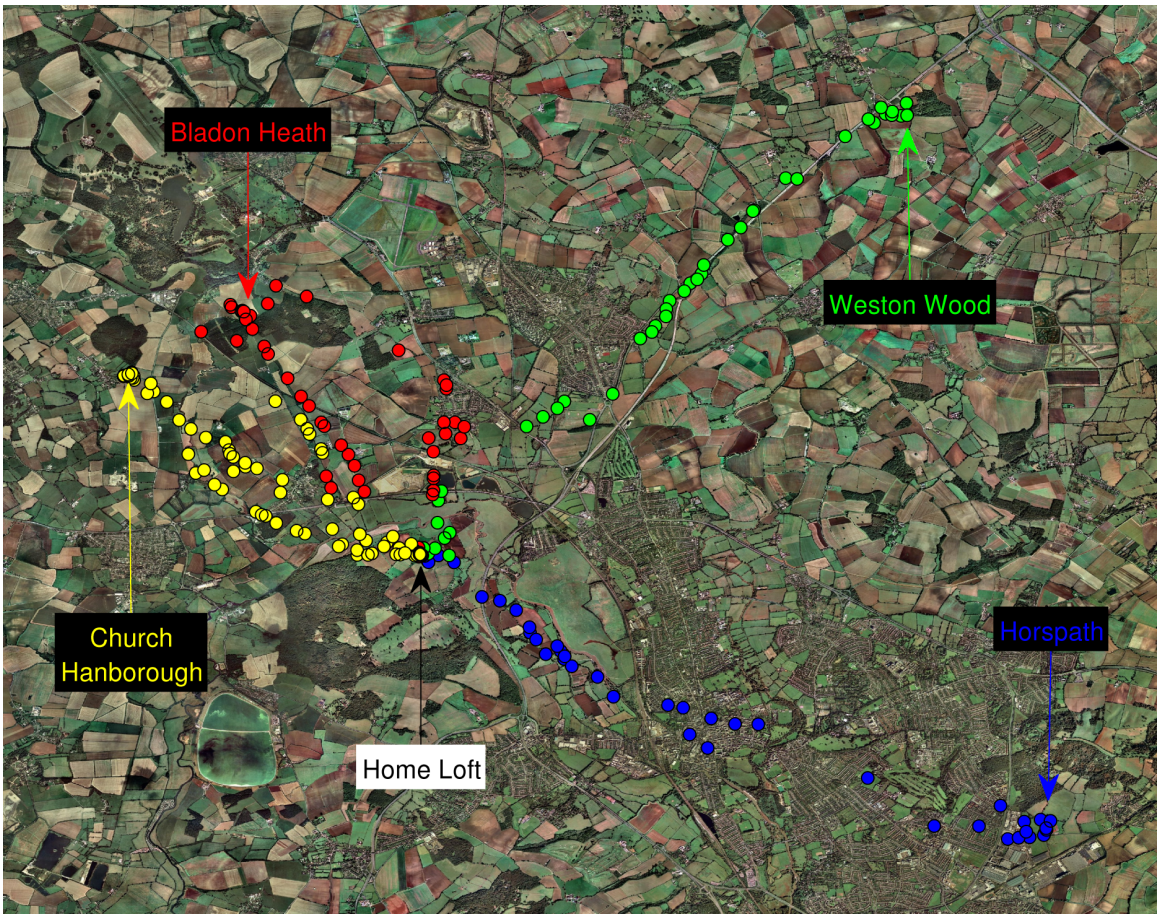


Figure 6.4: Landmarks identified from the four experimental release sites. Landmarks are colour-coded according to the release site they were identified from. The four release sites are labelled in the appropriate colour-coding, along with the home loft to which the pigeons returned.

sites.

6.8 Discussion

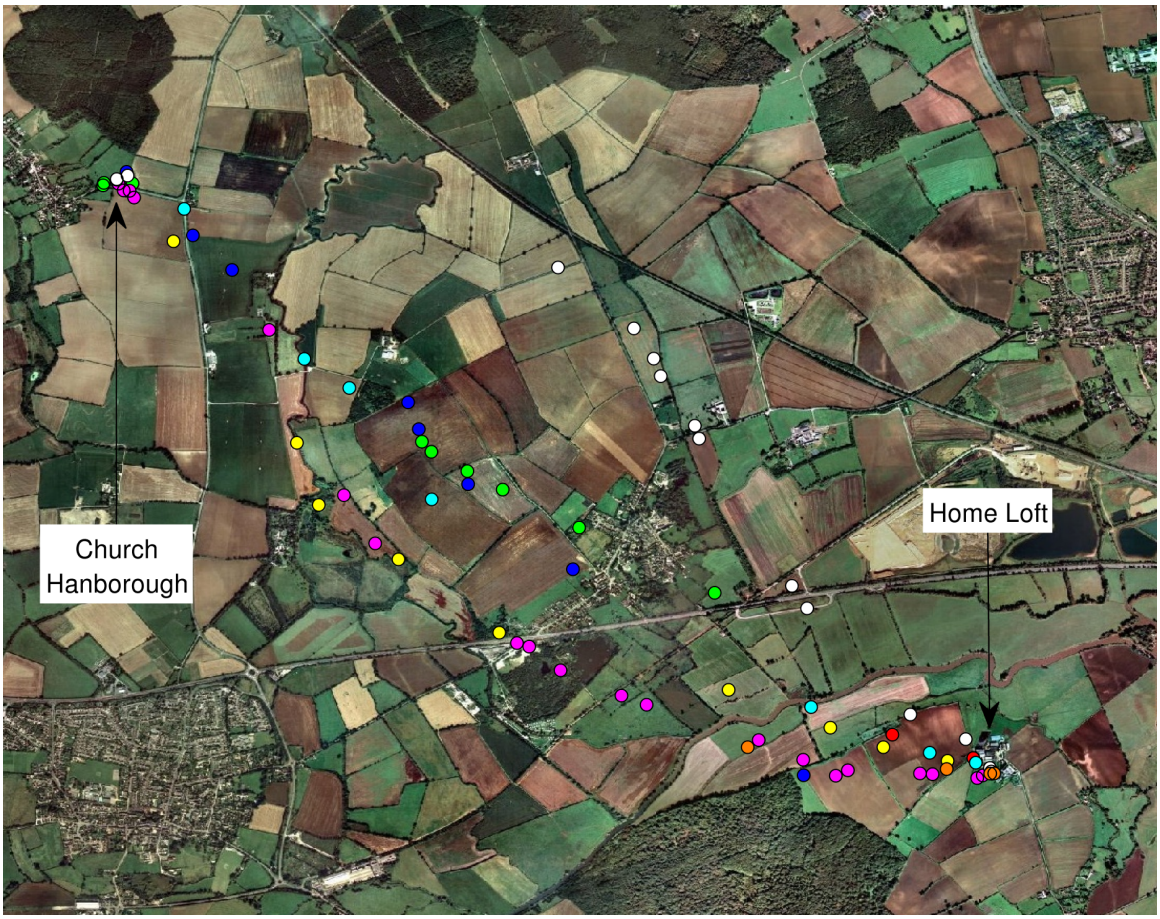
The optimal predictor method represents a genuinely novel approach to the problem of identifying important locations based on animal movements. The standard approach, using a combination of metrics derived from the movements to identify a 'signature' pattern associated with the points of interest suffers from a great problem, lack of sufficient training data. In the observational studies that constitute the great majority of the animal movement literature it is almost impossible to find a set of known

(a)

Figure 6.5: Identified landmarks from each release site, with each landmark colour-coded according to which bird it was identified from (note, the same colour indicates different birds for different sites). Release sites are labelled in each image: (a) Bladon Heath; (b) Church Hanborough; (c) Horspath; (d) Weston Wood. This panel: Bladon Heath

locations that exhibit the required properties. For example, in the study of pigeon flights it is not practical to experimentally adjust the landscape in a controlled manner or introduce known waypoints that one can be sure the birds will use. Without a set of such points to train an algorithm on it is impossible to learn the characteristic patterns associated with their use. The use of metrics to identify points therefore relies on the setting of arbitrary thresholds and the subjective judgement of the analyst.

By attempting to mimic the way a bird uses waypoints in this algorithm I hope to side-step this problem. If certain assumptions about waypoints are true then



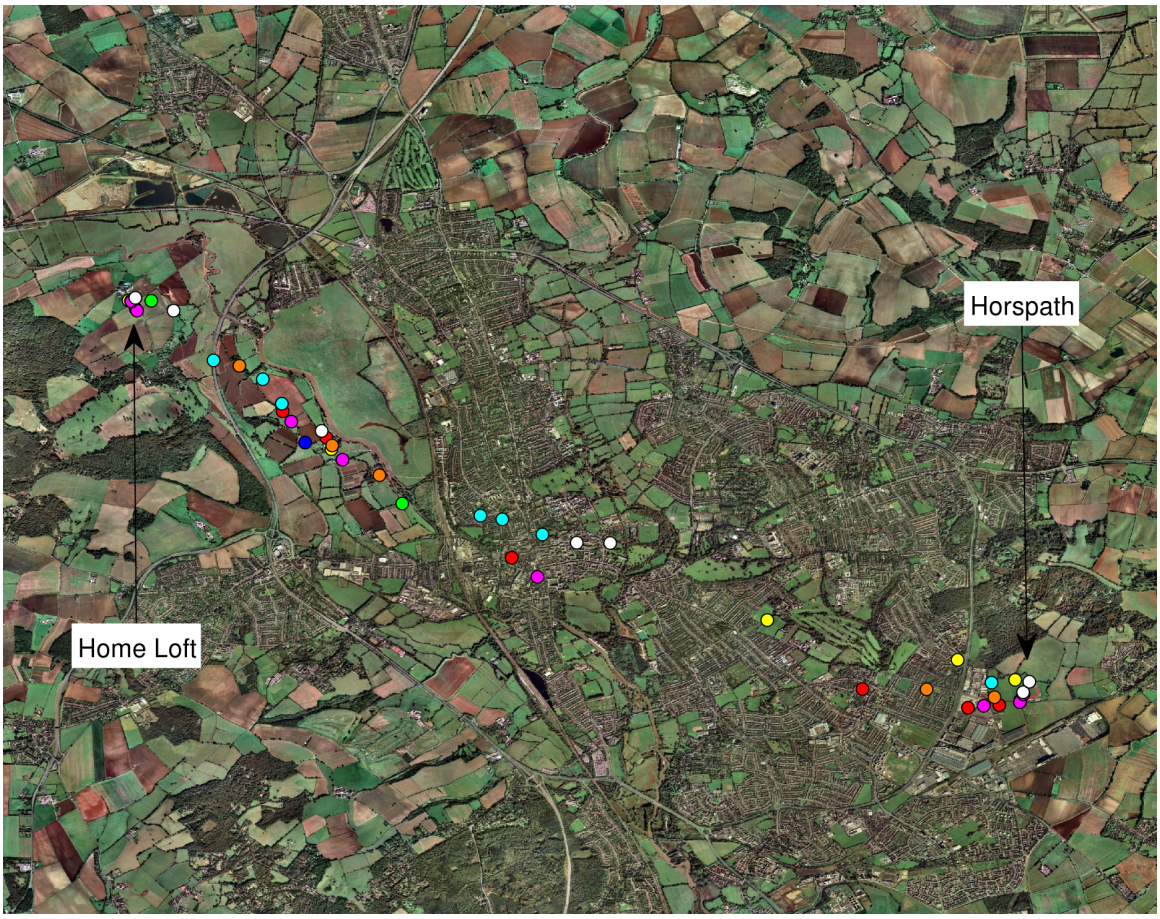
(b)

Figure 6.5: ...Continued. This panel: Church Hanborough.

this algorithm should identify them without requiring a training set to learn their characteristics. Those assumptions are:

- That birds primarily use waypoints to navigate in the familiar area
- That waypoints are geo-stationary locations that the bird physically visits during flight as opposed to distant landmarks

By further assuming that birds are maximally efficient in encoding their route it is possible to begin to estimate not only the locations of waypoints but also the correct number. However, it is likely that birds encode routes somewhat redundantly, since the cost of remembering an extra waypoint is surely small compared to the potential cost of getting lost. There is also a complex issue of calculating the correct number of

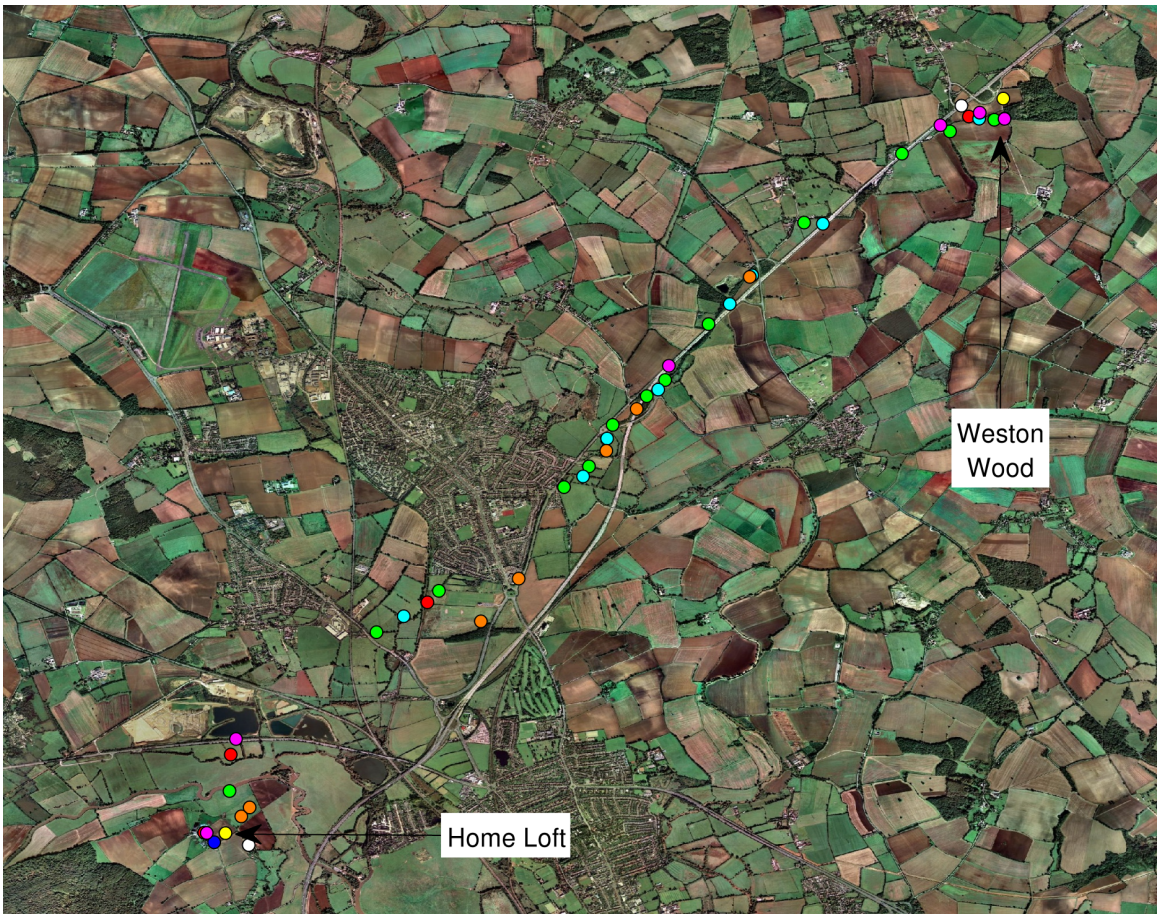


(c)

Figure 6.5: ...Continued. This panel: Horspath

waypoints to use in this model, since full Bayesian marginalisation over the waypoint locations is impossible for more than a very few waypoints.

Although I do not yet have an objectively determined link between waypoint locations and the underlying landscape, subjective visual inspection suggests a strong link between large scale landscape boundaries and waypoints. While some waypoints undoubtedly appear over small features, such as hedgerows or isolated buildings, the most consistently selected points across different birds are on the boundaries between environments, such as between fields and forest or between fields and urban areas. These large scale features appear far more important than the small scale information content of the landscape. Indeed, very few waypoints are discovered in consistently urban areas (such as from the Horspath release site flying over Oxford), despite the



(d)

Figure 6.5: ...Continued. This panel: Weston Wood

large amount of information potentially available to the bird.

The similarity of the pigeons' responses to forested areas and similarly sized urban areas suggests that they may be perceived equally by the bird, at least from a distance while in flight. Pigeons reared in a loft in a rural environment are likely to be unfamiliar with urban environments, unlike feral pigeons that thrive in cities. Therefore they may exhibit a hard-wired instinctive response to environments with a high degree of small-scale visual complexity — a property of both forests and towns.

Chapter 7

Collective Navigation and Modelling Pair-Released Flights

7.1 Introduction

This chapter explores how the distribution of individual flight paths can be extended, to model flight paths generated when pigeons are released as pairs. A simple Gaussian process mixture model is proposed, whereby the paired flight paths are considered to be generated by a combination of both the individual's own path distribution, and that of its partner. The evidence for various hypotheses regarding the nature of this mixing is calculated and compared. Having selected an optimal model, simulations of a previously performed experiment are made to show that important observed characteristics of the pair interaction are predictable from previously recorded individual flights.

7.2 Group Navigation

Animals come together for a huge variety of reasons, and derive a commensurate spectrum of benefits from remaining in groups. These include [Sumpter, 2008], among others:

- Reduced risk of predation. By combining closely in large groups, prey animals present collectively less 'surface area' to potential predators, and need to devote less time individually to watching out for predators — a job that can be confined

to a small number of individuals in very organised groups.

- Reduced energy consumption. Migrating birds can reduce the energy required during flight by adopting aerodynamic group formations, such as the classic 'V' formation, where each bird is subject to less air-resistance as a result of remaining close to the bird in front.

- Improved navigation. The many wrongs hypothesis [Bergman and Donner, 1964, Hamilton, 1967, Wallraff, 1978, Simons, 2004] suggests that groups are able to navigate more accurately by pooling information. If a group of animals effectively pool their estimates of the correct direction, the Central Limit Theorem (or 'law of large numbers', see MacKay [2003] or any statistics text) shows that a more accurate collective decision can be made.

It need not be assumed that all benefits of being in a group constitute 'reasons' for aggregating. It is perfectly possible, for example, that reduced predation may have been the primary selective pressure that led to the evolution of a aggregating instinct in so many animals, and that the additional consequences may be beneficial side effects.

The final item in the list above is the aspect of group behaviour under investigation in this chapter — how the navigation of birds is affected when they move collectively rather than as individuals alone.

As suggested above, if birds, or any other group of animals, are aiming to reach a common target, they ought to be able to improve their chance of success by effectively combining each individual's 'best guess' of the correct direction to move in. Under very loose assumptions about the distribution of the individual directional error the group error will scale as $O(N^{-1/2} \times$ average individual error) in a group of N identically distributed individuals. This theoretical benefit may break down under certain conditions however. It is difficult to say how well animals are able to communicate their directional estimates. It is most plausible they act according to simple, local behavioural rules rather than through global communication with the

group. An individual's best estimate of the direction may be apparent from its initial heading but there may be no effective communication of the uncertainty associated with that estimate. In such cases pooling information can be ineffective if individuals' estimates are not appropriately weighted. Even if an appropriate weighting is used, the effective group size may be dramatically lower if many members are naive and express large estimate variances, since these estimates will contribute little to the overall group decision.

Being in a group has costs as well as benefits. As shown in Sumpter [2008], the balance between costs and benefits with varying group size determines the optimal size of animal groups. An individual's navigational goals can potentially be detrimentally affected by moving with other individuals. Whilst averaging each individual's estimate of the same ultimate goal can lead to improvements in navigation, by accepting a group decision to move in a particular direction each individual runs the risk that other members of the group may in fact have different goals. By following the group towards the wrong target the individual may lose out by becoming lost (if it is taken outside of its familiar area) or by failing to achieve some aim associated with the target (e.g. foraging, feeding young, reaching mating grounds). Therefore it is in the individual's interest to be able to make an informed decision about whether to remain in the group.

For each individual, choosing whether or not to join or remain in a group, and weighing up the costs and benefits of doing so, can be seen as a model selection problem. This links back to the ideas discussed in Chapter 3, Section 3.4.2. Consider the following situation: A bird wants to navigate back to its home loft. Other birds nearby are moving as a group in approximately the correct direction, so the first bird joins the group, hoping to benefit from the improved group estimate of the homeward direction (as well as lower predation risk etc.). Later the group begins to move in a new direction, which disagrees with the first bird's estimate of the homeward direction. If the group decision sufficiently contradicts the individual's estimate of where it needs to go, it might infer that there is a conflict of goals. As in Section 3.4.2,

model selection can be performed by weighing the evidence in favour of competing hypotheses. A correctly reasoning Bayesian agent (which by no means is necessarily synonymous with a real animal) must select between alternative hypotheses, $H0$: 'my estimating facility is faulty/noisy' and $H1$: 'the group is estimating a different goal'. Model selection techniques can infer which of these hypotheses are more likely. To act on this judgement though, the individual must weigh the relative likelihoods of these hypotheses by appropriate loss functions representing the cost of acting alone and the cost of aiming for the wrong goal. This way the individual can maximise its expected gain or *utility*.

The approximation of such reasoning in real animals produces conflict between the flocking instinct and individual preferences. Of course, it is unlikely that animals genuinely reason as above. More likely is that selective pressures have favoured behavioural patterns that approximate optimal decision making. Understanding how the conflict between individual and group preferences is resolved lies at the heart of studies of group navigation. How animals communicate their preferences and how information propagates determines how a group will behave, whether they will improve their navigational outcomes (efficiency, successfully reaching the target, etc.) and whether the group will remain cohesive or divide into sub-groups with more aligned preferences.

Models of collective movement have demonstrated that this complex trade-off between individual preferences and collective cohesion can emerge from simple behavioural rules. Couzin et al. [2005] show that agents that apply only very simple, localised behavioural rules can form groups that exemplify this 'decision' problem (the group in this case does not actually decide to take any action, the actions emerge from the underlying rules). Each agent is programmed to avoid other agents that are too close, and to align with other agents within a sphere of influence around the individual. In addition, some agents are given a preferred direction of travel, corresponding to partially informed individuals who have differing estimates of where their goal lies. When these differing preferences are sufficiently similar the group remains cohesive,

and the group as a whole moves in a directional average of their individual preferences, as predicted by the Many Wrongs hypothesis. However, when the preferences of the informed individuals become to divergent the group begins to fracture. When two subgroups of informed individuals exist, with a large discrepancy between the subgroup preferences, the group splits into two — the differences between the individual preferences are too large to sustain group cohesion.

Applying this theoretical work to experiments on collective avian navigation, recent work by Biro et al. [2006b] has focused on a small system of two interacting homing pigeons as the basis for larger scale systems. The researchers experimentally produced a conflict between the two individual birds in each pair and observed how the conflict was resolved as a function of the degree of conflict. The results of that paper support both the theoretical results from large groups, such as in Couzin et al. [2005], as well as a supporting model for the two-bird system that accompanied the experimental results. This system forms the basis of the work in this chapter, which aims to explain the experimental results of Biro et al. [2006b] by simulating experiments using the Gaussian process model developed in the earlier chapters of this thesis.

7.3 Concept

Biro et al. [2006b] performed an experiment where pigeons, all individually familiar with a release site, were released as pairs. The results showed that the birds displayed three distinct modes of behaviour. Some pairs split up and flew along their previously established habitual routes. Other pairs compromised over route selection, flying along an average of the two habitual routes. In the third case one bird acted as the leader, as both birds flew along the leaders habitual route. It was discovered that the type of behaviour observed was strongly dependent on the divergence of the birds' habitual routes. When the habitual routes were sufficiently close, birds tended to compromise. As the divergence grew beyond a critical threshold the birds

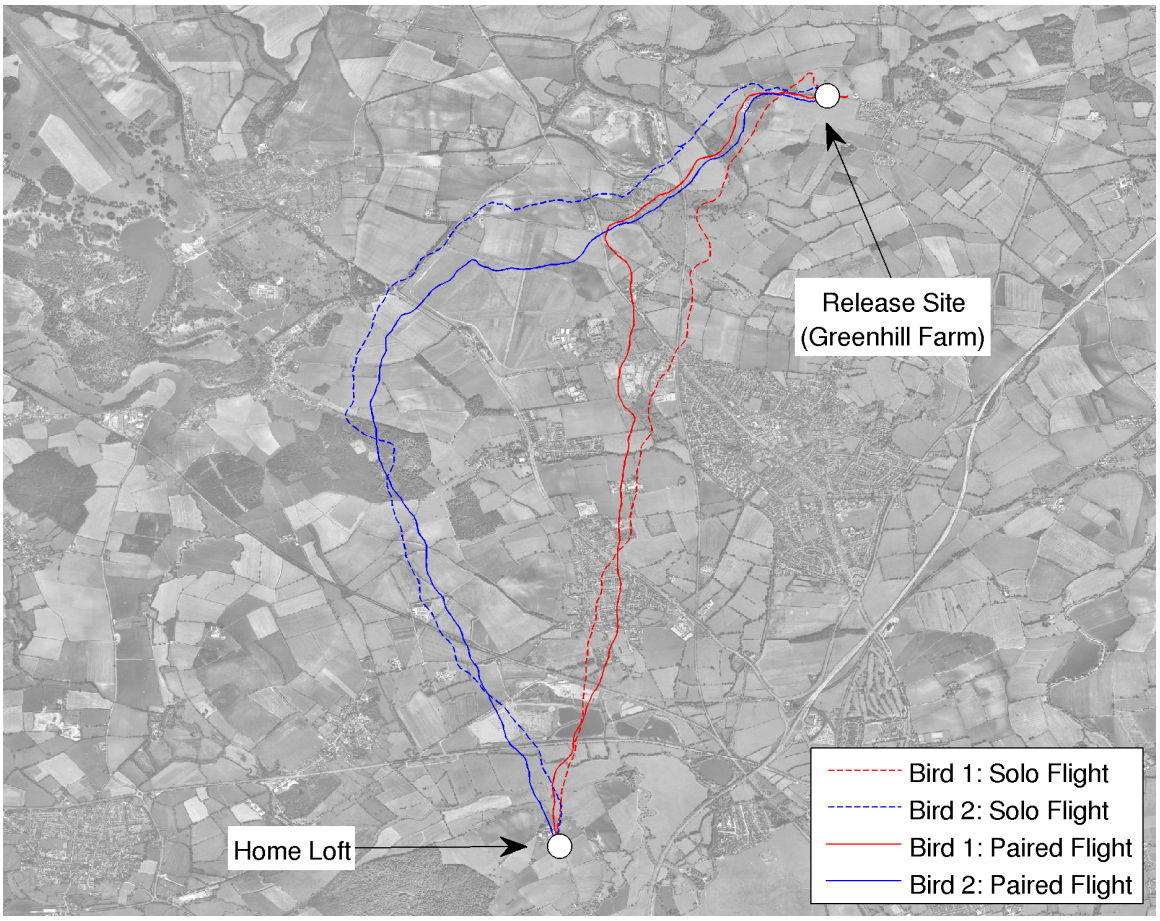


Figure 7.1: An example of a paired-release experiment. The dashed lines show the final solo flight of each bird prior to the paired release. The solid lines show the flight paths of each bird when released together. The birds leave the release site together and remain together until their habitual paths become too divergent; once they split up they remained apart.

tended to switch, either splitting up or transitioning to a leader-follower relationship. The transition between the compromise behaviour and the leadership or splitting behaviour was empirically observed by inspection of the distribution of each bird's distance from its own established route (as defined by the solo flight). The distribution of this distance as a function of the distance between the two established routes reveals an emerging bi-modality when the established routes differ by more than a critical distance. To determine exactly where this bi-modality arises, and therefore what the critical separation is, Biro et al. [2006b] used the kurtosis of the distribution as an indicator of bi-modality.

Figure 7.2 is a reproduction from that paper. Plot (a) shows the distribution of the distance between the paired flight paths and the respective established route. Plot (b) indicates the two maxima of this distribution with increasing separation of the established routes. For small separations the distribution is uni-modal, reflected in the fact that the two maxima are close together. For greater separations two distinct peaks emerge and the maxima are widely separated, indicating that the birds are either on their own established route, or following the partner's established route. Plot (c) shows the kurtosis of that distribution, indicating the critical separation of the established routes at which compromise behaviour stops and leadership begins. Given the confidence intervals on when the kurtosis becomes negative, the critical separation is in the range of 300-600 metres.

Biro et al. [2006b] also found paired flight paths were, in general, more efficient than the solo flights. This was unsurprising in the case of the birds that were least efficient originally, since the other bird could be expected to 'pull' them towards its own, more efficient route. In cases where the established routes were on opposite sides of the straight route from release site to home loft an improvement in both birds could be expected, since both would be 'pulled', by their partner, towards the centre. However, the surprising result was that, on average, the more efficient bird from each pair also improved in efficiency during the paired flight, even when both established routes lay on the same side of the straight line. Figure 7.2 (d) shows the improvements in efficiency in each pair, plotted as the improvement by the previous most efficient bird versus the improvement in the previous least efficient bird. In the large majority of cases the least efficient bird of each pair improved. In fewer cases, but still the majority, the previously more efficient bird also improved.

Thirdly, and most surprisingly, it was found that the birds in the experiment obeyed a 'transitive dominance hierarchy' with respect to leadership in the pair, i.e. if bird A followed bird B, and bird B followed bird C, then bird A would follow bird C if they were released as a pair. This was striking as the completely transitivity suggested that a very well determined factor was involved in the decision to follow

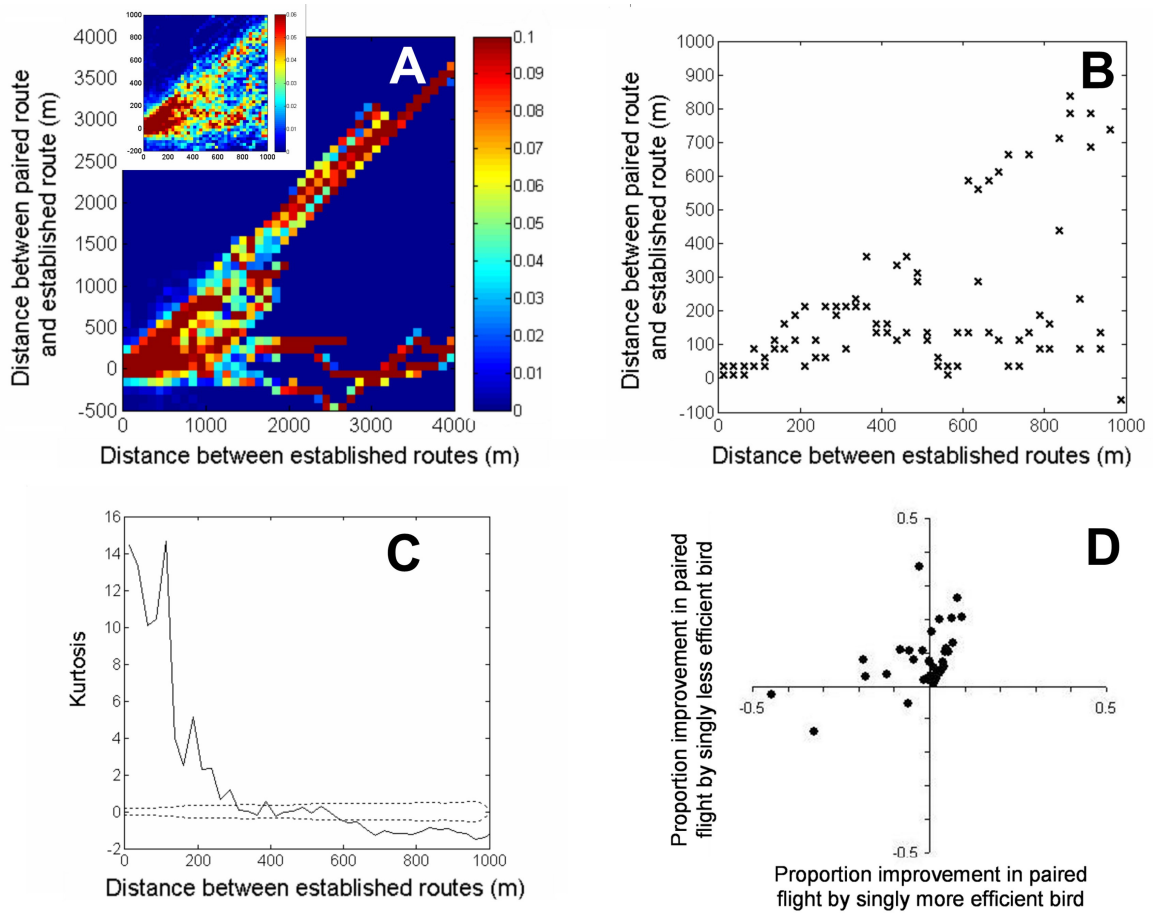


Figure 7.2: **From Biro et al. [2006b], reproduced with permission**. Original caption: Analysis of routes taken by pigeons released in pairs as a function of the distance between the birds respective established routes. Data from pairs where birds split up have been excluded. (A) Distance of birds from their own established routes during paired flights as a function of the distance between their own and their flight partner's established routes at the corresponding stage of the journey. Positive numbers on the y-axis indicate positions assumed by the birds that lay in the direction of the flight partner's established route (i.e., within the area enclosed by the two established routes or on the far side of the partner's route); negative numbers correspond to positions in the opposite direction (i.e., away from the partner's route). Inset magnifies the 0-1000 m range. (B) The two modes of the data shown in (A) inset. (C) Kurtosis of the data shown in (A) inset. Dotted lines correspond to the upper and lower boundaries of the 95% confidence interval for kurtosis consistent with a Normal distribution; significant bi-modality in the distribution of the data begins to emerge when kurtosis drops below the lower boundary. (D) Proportion improvement during paired flight by the bird with the less efficient (longer) established route as a function of proportion improvement by the more efficient bird of the pair. Proportion improvement was calculated as the difference between track length during paired flight and established route length, divided by established route length. Negative values correspond to loss of efficiency during paired flight compared to the same individual's established route.

or lead, yet the hierarchy was unconnected to potential predictors such as previously established route efficiency or social dominance ('pecking order') at the loft.

Biro et al. [2006b] supported their findings with reference to a simple mathematical model that accompanied the experimental results. The essence of the model was that each bird was subject to two competing forces. Firstly, each bird was attracted towards a different stationary point, which represented attraction to a previously established habitual route. Secondly, the birds were mutually attracted towards each other. These attractions were assumed to initially grow in strength with separation, but eventually decline in strength once the separation was sufficiently great — to represent losing contact with either the habitual route or the other bird. Similar to the model of Couzin et al. [2005], this model predicted that the birds would remain together, half way between their respective stationary attraction points, up to a critical separation threshold. Beyond this threshold the birds would either return to their own attractor, or follow the partner towards the other attractor.

This prediction was supported by the experimental findings, and subsequent work by Sumpter et al. [2008] showed that differences in the degree of attraction towards the stationary points could potentially explain the existence of the transitive navigational hierarchy, since small differences led to highly predictable leadership. However, some questions remained unanswered. Firstly, why does the critical separation threshold take the value it does? The original model can not predict the value of this threshold, since it has a number of adjustable parameters that must be tuned to fit the experimental results, including some parameters that can only be learnt by observing paired behaviour. Secondly, if differing attraction strength is the cause of the observed leadership hierarchy, can these strengths, or some proxy, be observed, such that the leadership hierarchy could be predicted without reference to actual paired flights?

In this chapter I use the previously developed Gaussian process model to address these questions. I create a Gaussian process mixture model that will remove some of the complexity of the individual based models of the original work, by ignoring the

details of the bird-to-bird interaction. Instead I will assume that each bird is simply influenced by its own habitual route, and the habitual route of its partner. Of course, in reality, each bird can not 'see' its partner's habitual route. However, I will assume that the effect of the bird-to-bird interaction effectively provides this information, mediated through the partner's movements.

7.4 Data

Track data were recorded using established protocols [Biro, 2002, Biro et al., 2002], similarly to the data used in Chapter 5. Three release sites were used: Greenhill Farm (distance 8.6km, bearing to home 197 degrees), Weston Wood (distance to home 10.6km; direction to home 221 degrees) and Church Hanborough (distance to home 5.3km, direction to home 129 degrees). The positions of Weston Wood and Church Hanborough can be seen in Figure 5.1; the position of Greenhill Farm is indicated in Figure 7.1.

A total of 22 individual birds were used in the study. Tweleve were assigned to Greenhill Farm, 6 to Weston Wood and 4 to Church Hanborough. Each bird was given twenty solo training flights from the assigned site.

Following training, birds were randomly assigned into pairs and simultaneously released from their respective training sites. On subsequent releases, birds were again randomly assigned, but only to novel partners. Between paired flights, birds were individually released at least twice and, provided that they were still recapitulating, assigned to their new partner. Birds that did not demonstrate satisfactory route recapitulation after a paired release did not participate any further. Most birds participated in three to six pairs; one completed two paired flights, and another completed seven. A total of 48 paired releases were performed across the three sites.

In each experiment two individuals were chosen. They first made one tracked solo flight, followed by a tracked flight as a pair (a total of four tracks per experiment). Each solo flight was checked to ensure the bird recapitulated its previous habitual

route. If it did not the experiment was aborted.

Further details of experimental procedure can be found in Biro et al. [2006b].

7.5 Model Extension and Selection

Equation (54) gives the probability distribution for future flights by an individual, based on previously observed flights. Call this the solo flight model, M_{solo}, and denote this distribution as $P(\, x \mid X_i, M_{\text{solo}}\,)$, where X_i represents the set of previous solo flight path observations. Consider the following simple model of pair interaction. A mixture model of both birds' individual flight path distributions represents the probability for a flight path while in the pair. Denoting this model as M_{pair}, the probability of observing a given path during paired flight is

$$p(\, x_{pi} \mid X_i, X_j, \alpha, M_{pair}\,) = \alpha p(\, x_{pi} \mid X_i, M_{solo}\,) + (1-\alpha) p(\, x_{pi} \mid X_j, M_{solo}\,)\,, \quad (58)$$

where x_{pi} represents the flight path of bird i while paired with bird j, and $\alpha \in [0,1]$ is a mixing coefficient that determines how strongly the bird remains attached to its own habitual route and how much the partner bird influences it. A number of varying hypotheses are available to determine how this interaction is modulated, and thus how the mixing coefficient is determined:

- M_1: The null hypothesis. Birds do not interact and their flight probabilities remain unchanged, therefore

$$p(\, x_{pi} \mid X_i, X_j, M_{\text{pair}}\,) = p(\, x_{pi} \mid X_i, M_{\text{solo}}\,) \quad (59)$$

- M_2: The mixing coefficient is a release-specific quantity. In any given paired release the ratio of attraction to the partner's route and fidelity to one's own route is essentially generated at random, therefore

$$p(\, x_{pi} \mid X_i, X_j, \alpha, M_{\text{pair}}\,) = \alpha p(\, x_{pi} \mid X_i, M_{\text{solo}}\,) + (1-\alpha) p(\, x_{pi} \mid X_j, M_{\text{solo}}\,) \quad (60)$$

- M_3: The mixing coefficient, now denoted as α_i is a property of the individual, bird i. A bird which is loyal to its own route in one release will be loyal to its

own route consistently, regardless of the partner, giving as the probability for a path

$$p(\, x_{pi} \mid X_i, X_j, \alpha_i, M_{\text{pair}}\,) = \alpha_i p(\, x_{pi} \mid X_i, M_{\text{solo}}\,) + (1-\alpha_i) p(\, x_{pi} \mid X_j, M_{\text{solo}}\,) \tag{61}$$

- M_4: Instead of being a property of the modelled individual the mixing coefficient, denoted as $\beta_j \in [0,1]$, is determined by the partner, bird j. A bird which successively attracts its partner in one release will do so consistently, regardless of the identity of the other individual. The probability of the path is therefore

$$p(\, x_{pi} \mid X_i, X_j, \beta_j, M_{\text{pair}}\,) = (1-\beta_j) p(\, x_{pi} \mid X_i, M_{\text{solo}}\,) + \beta_j p(\, x_{pi} \mid X_j, M_{\text{solo}}\,) \tag{62}$$

- M_5: The mixing coefficient emerges from the interplay of both individuals. Both route loyalty and cross attraction contribute. There are now two mixing coefficients, denoted as α_i, for the route loyalty of bird i, and β_j, for the cross attraction of bird j. Therefore the probability is given by

$$p(\, x_{pi} \mid X_i, X_j, \alpha_i, \beta_j, M_{\text{pair}}\,) = \frac{\alpha_i}{\alpha_i+\beta_j} p(\, x_{pi} \mid X_i, M_{\text{solo}}\,) + \frac{\beta_j}{\alpha_i+\beta_j} p(\, x_{pi} \mid X_j, M_{\text{solo}}\,) \tag{63}$$

The marginal likelihood, or *evidence*, of each of these alternatives can be calculated by appropriately marginalising over the possible values of the mixing coefficients. The Bayes factor is the correct way to select the best model (see Chapter 3). The matrix of Bayes factors is defined as:

$$\text{BF}_{ij} = \frac{\prod_{\text{Experiments}} \prod_{k=1}^{2} p(\, \text{x}_{\text{pk}} \mid M_i\,)}{\prod_{\text{Experiments}} \prod_{k=1}^{2} p(\, \text{x}_{\text{pk}} \mid M_j\,)}. \tag{64}$$

Using uniform prior distributions on all parameters to calculate the marginal likelihood of each model, we obtain the full matrix of Bayes factors, **BF**, with each element calculated according to equation (64). This matrix, expressed in below in logarithmic form, shows the pairwise comparisons between model M_i along the first axis (vertical)

and model M_j along the second (horizontal). Positive numbers indicate that M_i is favoured over M_j (i.e. the marginal likelihood of M_i is higher). The diagonal elements are identically zero since these express the Bayes Factor between two identical models.

$$\log_2 \mathbf{BF} \,/\, \text{bits} = \begin{bmatrix} 0 & -729.1 & -736.0 & -729.3 & -735.3 \\ 729.1 & 0 & -6.9 & -0.2 & -6.2 \\ 736.0 & 6.9 & 0 & 6.7 & 0.6 \\ 729.3 & 0.2 & -6.7 & 0 & -6.0 \\ 735.3 & 6.2 & -0.6 & 6.0 & 0 \end{bmatrix}. \tag{65}$$

These numbers show a strong re-affirmation that paired birds do interact (rejection of M_1). Moreover they suggest that the most important factor in this interaction is the loyalty of an individual to its own route. M_3 emerges as the strongest model, implying that a consistent loyalty to the habitual route is a better predictor of behaviour than a consistent tendency to attract the partner. There is, however, little evidence to separate M_3 and M_5, so the attractiveness of the partner can not be dismissed as a possible influence.

Using the selected model, M_3, simulated re-creations of the experiments carried out by Biro *et al* [Biro et al., 2006b] were performed. The standard GP model of Chapter 5 was used to learn the individual distributions of the experimental birds, based on the last 5 training releases of each bird. Simulations were then made of the paired releases carried out in the genuine experiments, choosing a random value of α_i uniformly between 0 and 1 for each bird. Simulations consisted of sampling from the distribution defined by Equation (61), given the randomly selected values of α_i for each bird. The complete set of experiments was repeated 10,000 times to obtain a wide sampling of possible outcomes. The simulation results were then analysed, using the methodology of Biro et al. [2006b]. The same simulations were also performed using model M_2 for comparison; this model best expresses ignorance about the coupling between the birds, so results that are robust to switching to this model can be deemed predictable without any observation of paired flight.

7.6 Simulation Results

Analysis of simulated experiments show that the transition between compromise behaviour and leadership/splitting behaviour is faithfully reproduced. Repeatedly simulating the true pairings produces a predicted result of the kurtosis analysis as shown in Figure 7.3. The solid line indicates the median result of the simulations, while the dashed lines indicate the bounds of the inter-quartile range. This closely matches the trend seen in Figure 7.2 (c). The range of possible experimental values for the critical separation is shown by the shaded area. This is determined by the interval during which the experimental kurtosis is within the 95% confidence interval indicated in that figure. In Biro et al. [2006b] the critical separation was taken to be the greatest value within this interval, when kurtosis was definitively below zero (below zero and outside of the confidence interval). The equivalent value was calculated on each of the ten thousand simulations. The distribution of critical separation values in those simulations is shown in Figure 7.4. The value of the experimental critical separation is also indicated by the dashed line. The experimental value falls within one standard deviation of the mean of the simulation results, and is thus strongly consistent with the simulation predictions. In both Figure 7.3 and Figure 7.4 the experimental value of the critical separation is slightly below the prediction from simulations. While the results are still consistent, this divergence could be due to an excessive 'prior ignorance' about the likely mixing coefficients in Equation 61. The simulations are performed by sampling these coefficients from a uniform distribution. However, it is reasonable to believe that a bird's attraction to its own route is generally greater than its attraction to the other bird's route — an effect which has to be mediated through attraction to the other bird. Repeating the simulations with a more skewed prior distribution on these parameters results in a more precise prediction of the critical separation; the extent of the birds' preference for their own routes can be adjusted for an optimal fit. However, one key strength of these results is that the critical separation can be estimated to within a small error even under complete

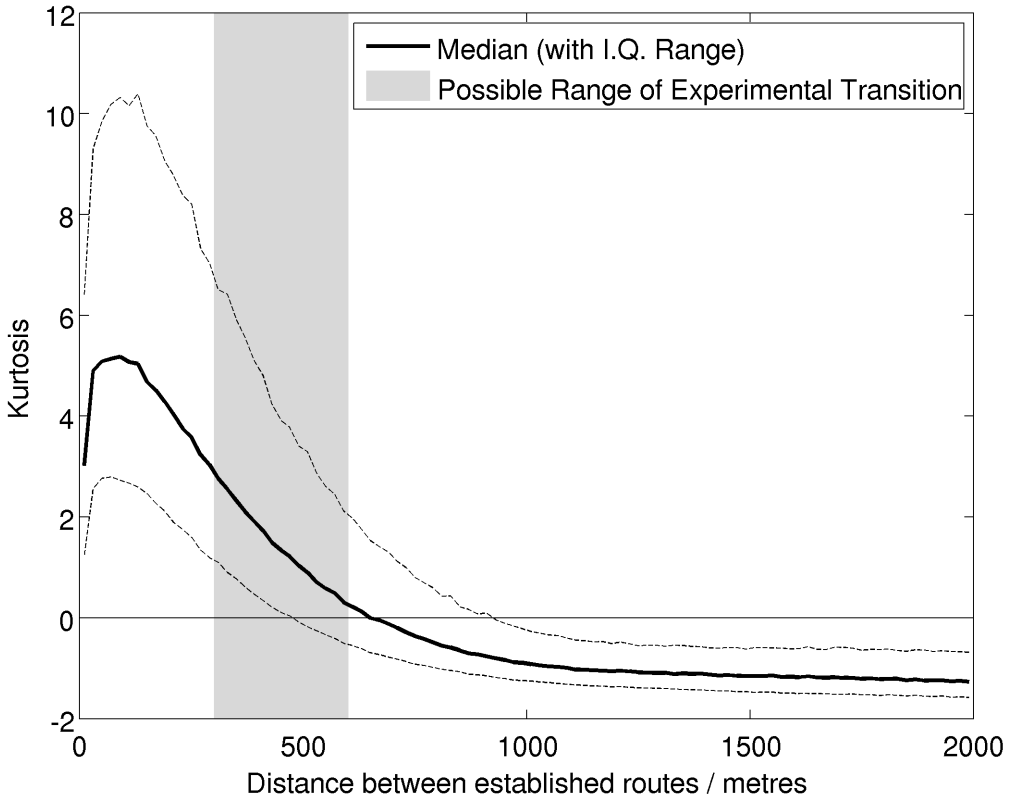


Figure 7.3: Kurtosis of the distribution of distance between paired flight and solo flight. This plot is the result of repeating the analysis shown in Figure 7.2 (c) on simulated data. The heavy line indicates the median value of the kurtosis averaged over the 10,000 simulations, with the inter-quartile range indicated by the dashed lines. The shaded region represents the range of values in Figure 7.2 (c) where the kurtosis is within the 95% confidence interval of zero, which gives a range of possible values for the critical separation value.

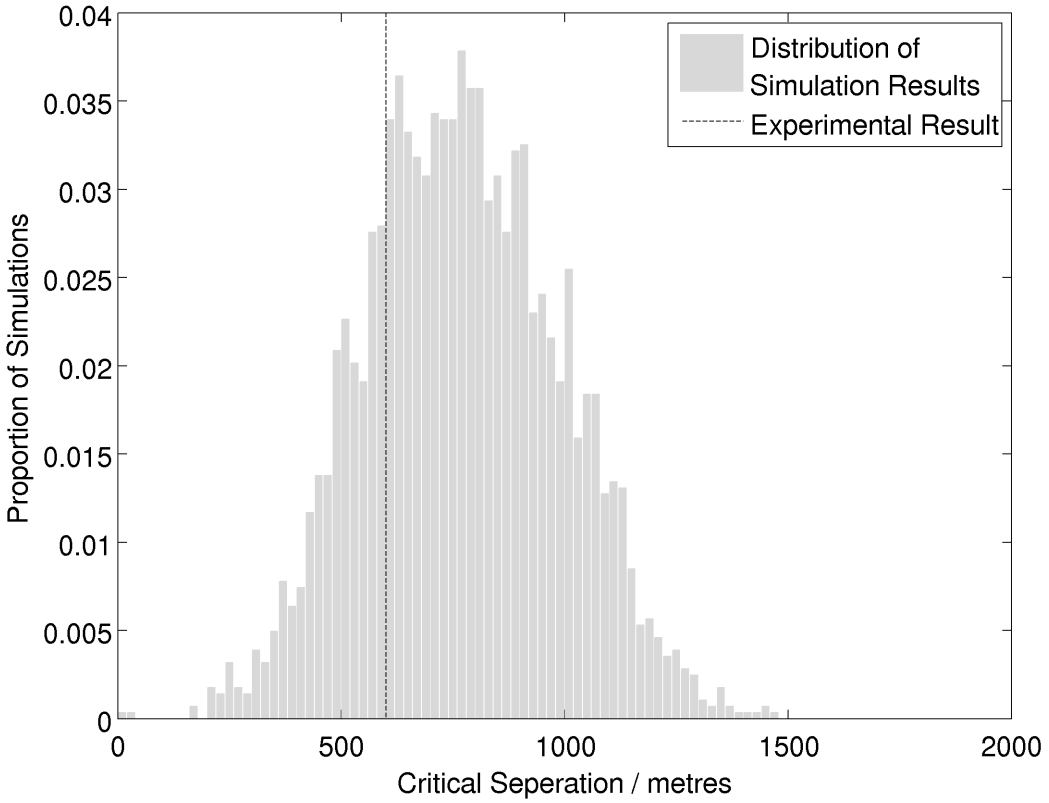


Figure 7.4: Distribution of the Critical Separation Value from Simulations. The distribution is derived from the separation value at which the kurtosis is first significantly below zero on each of the 10,000 experiments. The dashed line indicates the equivalent value from the experimental data.

ignorance of the coupling strength; therefore the results presented are those from simulations with uniform sampling distributions. Similarly, these simulation results are unchanged when the simulating model is changed from M_3 to M_2, indicating that these predictions can be confidently made without reference to observations of paired experiments.

By performing simulations based on the GP model I am able to make a successful quantitative prediction of the transition point based only on observation of individual behaviour. The model also provides a surprisingly simple means to visualise how divergent goals lead to compromise or leadership. Each bird as an individual effectively flies along a 'flight corridor' – an ideal habitual route surrounded by a Gaussian region of variation. The model represents the paired flight distribution as a mixture of the individual distributions. The sum of two Gaussian functions may be either uni-modal or bi-modal, depending on the separation of the means (the distance between the habitual routes) and the standard deviations (the width of the 'flight corridor'). The model would imply that birds compromise in regions where the combination of their individual distributions is uni-modal. Once the distribution becomes bi-modal they are forced to choose one of the established routes to remain in a region of high probability. Figure 7.5 demonstrates the transition from uni-modality to bi-modality is the sum of two one-dimensional Gaussian functions. Inspection of the true flight paths from each experimental release in relation to the previously established distribution of the individual flight paths reveals that this switch from uni-modality to bi-modality does not perfectly predict the exact location at which the birds will switch their behaviour for every experiment. Rather, it appears that the emergence of bi-modality in the distribution indicates the emergence of an untenable conflict. The exact point at which this conflict is resolved is currently unpredictable for a single experiment but over several experiments the average behaviour confirms to the model's predictions.

The simulations of the GP model cannot explain the navigational hierarchy. Because the mixing coefficients in the model are free parameters they cannot be known in advance of observing the paired flights. When performing simulations based on

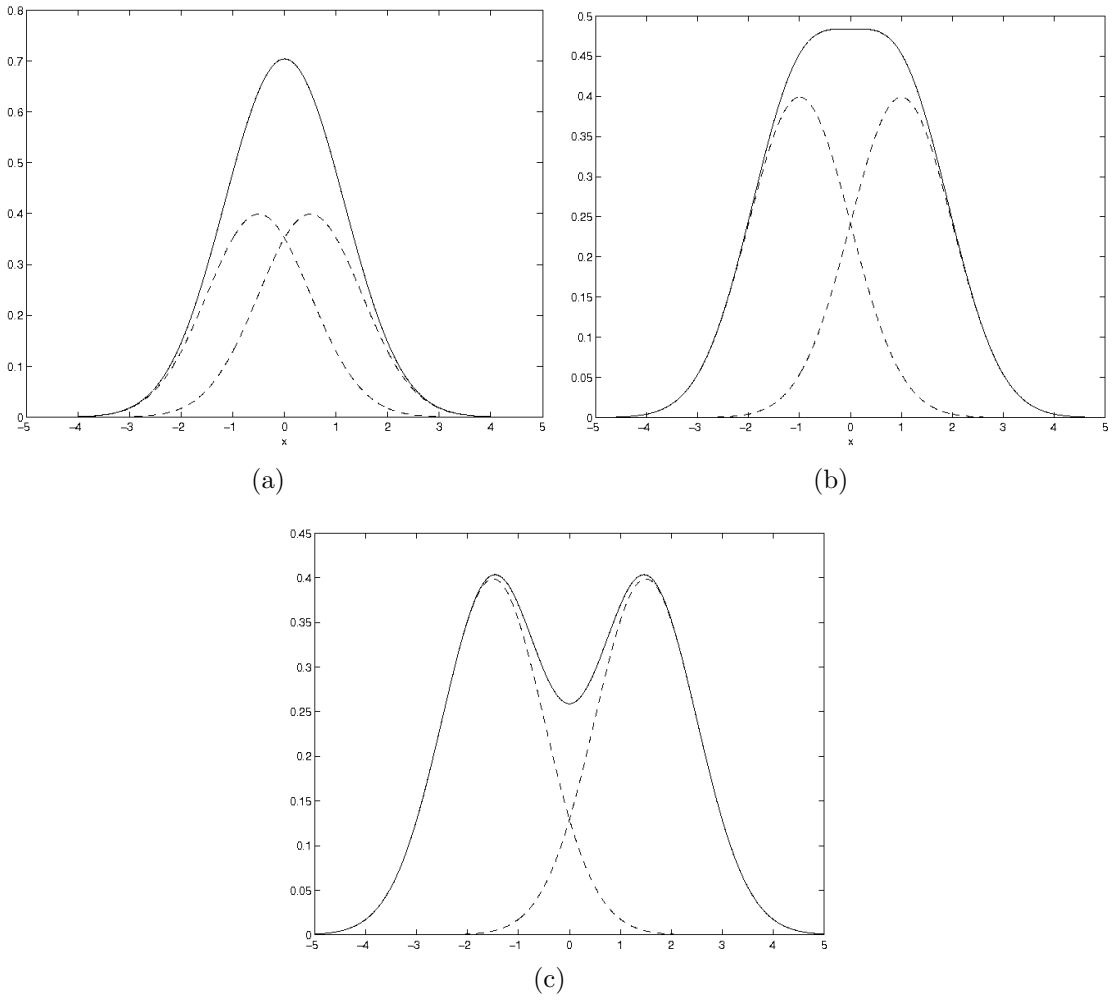


Figure 7.5: Uni-modality and Bi-modality in a sum of two Gaussian distributions. In plot (a) the two Gaussian functions (dashed lines) have centres that are narrowly separated in proportion to their widths. In this situation the sum of the two functions (solid line) is uni-modal. As the separation of the centres is increased the sum of the two functions passes through a transition point (plot (b)), before becoming bi-modal (plot (c)).

uniform sampling of these parameters each bird necessarily 'wins' a roughly equal proportion of the time. However, the GP model does provide an indication of how the hierarchy is constructed outside of simulation studies. The paired flight model is constructed as a mixture of individual flight path distributions. Each of these individual flight path distributions has an associated variance that defines the width of the 'flight corridor' for that individual bird (see the widths of the indicated flight corridors in Figure 5.4 for examples). Of the 35 experiments where the birds remained together throughout the entire flight, in 26 cases the bird with lower individual variance remained closest to their own established route — the definition of the leader of the pair. Hence there appears to be a correlation between the relative flight path variances of the two birds in each experiment, and the leader-follower roles they adopt. Here flight path variance may be a proxy indicator of the bird's attachment to its own route. Sumpter et al. [2008] showed that under the original model of Biro et al. [2006b], differing levels of attraction to the established route between the two birds led to very predictable leadership decisions. If greater attraction to the established route in general leads to lower individual flight variance this would explain why lower variance is associated with leadership.

7.7 Discussion

The flight paths of pigeons that are released in pairs show a complex interaction of different effects. Birds that have developed habitual routes that they would ordinarily follow with high fidelity are in many cases drawn away from these routes by the desire to remain close to the partner bird. Considering how predictable the individual flight distribution becomes under sufficient training (as shown in Chapter 5), the attraction to other pigeons must be a very strong and fundamental instinct to draw the bird outside of the strict bounds of its previous behaviour.

The interaction of two pigeons released together is reflected in the enormously greater probability of the observed flight paths under a model that incorporates the

pairing effect than under a null model with no interaction term. The effect on the flight path of flying with a partner is observed on multiple scales. On the broadest scale the bird may abandon its habitual route, or fly along an average of its own route and its partner's. On smaller scales, the paired flight paths appear to be more efficient because they are smooth and less tortuous over short distances, rather than because they necessarily lie closer to the straight line route overall. The GP mixture model is a simple way to simulate and visualise the large scale effects of paired flight. The spatial overlapping of Gaussian 'flight corridors' provides a large scale view of the degree of conflict between the birds as a result of their differing habitual routes, and in simulations predicts (within an acceptable margin of error) the point at which that conflict can no longer be resolved through compromise, and must instead result in either the birds splitting up or one bird abandoning its own route in order to remain close to the other bird.

The GP mixture model does not describe the small scale effects of paired flight. Within the simulations, birds that 'stay together' are further apart on average than in reality, since they are not truly interacting, except through having chosen the same individual habitual route to follow. All the inter-bird interaction in the mixture model is mediated through this one-off choice over which route to follow. Therefore the localised effects of interacting with another bird are not simulated. The GP mixture model only provides a macroscopic view of conflict and co-operation.

The variability of an individual's solo flight paths may have a role in determining how the conflict in navigational aims is resolved. In cases where birds remain together there is a trend for the bird with lower individual variability, as defined by the variance of their individual flight distribution, to become the leader — to remain closer to its own habitual route. It is not clear whether this trend is mechanistic – whether that bird happens to be closer to its own route at the moment the leadership decision is made, and thus more likely to remain on their own route – or whether the variability is a proxy for some quality the bird communicates to its partner – some intrinsic level of 'certainty' about its route, or a navigational ability. Indeed, it is not clear

in general if lower variability indicates that the bird is more skilled, or confident – a demonstration of the bird's skill at recalling a route – or less skilled and/or confident — an expression of the bird's unwillingness to leave the strict confines of the previous established route, potentially because of fear of getting lost, or because the route has so far proved safe from predation. Certainly route variability is not the definitive predictor of leadership, since the observed leadership hierarchy exhibits no deviations from complete transitivity, while route variability can only predict approximately three in four cases.

In conclusion, the very local interaction of two birds in flight, balancing their desire to remain close together against their instinct to remain on their learnt habitual routes, produces large scale effects that are predictable even in the absence of detailed knowledge of how the local interaction operates, much like the thermodynamic behaviour of gases can be described without a model of the microscopic structure of a gas. This may provide a methodology for understanding larger groups, describing them as a large mixture of competing preference distributions.

Chapter 8

Conclusion

Each of the preceding chapters detailing original work in this thesis has a section discussing the results and implications of that work. Therefore I want to use this chapter to discuss some of the broader issues relating to this work and its relationship to other work in the field of animal movement and navigation. I also want to take the chance to give further details on the current problems and limitations that bound the scope of the methodology as it stands and provide potential avenues for further development.

The emerging collaboration between the study of animal behaviour and the mathematical sciences has remained somewhere short of a full marriage of ideas. On the one hand, many zoologists have applied pre-developed, 'off the shelf' algorithms to situations that closely resemble previously addressed mathematical problems, for example, in the use of the Kalman filter and other noise reduction algorithms to improve the accuracy of noisy tracking technology. On the other hand, mathematicians and physicists have created novel theoretical models of animal movement, based on the same principles used to model the kinematics of gases [Vicsek et al., 1995], which has led to an entire class of models of perfect agents that respond predictably to the other agents in their vicinity according to simple rules. These models have proved remarkably adept at simulating the qualitative forms of behaviour seen in flocks of birds, schools of fish and swarms of insects, but rarely do these models interact with real observed data. When an attempt is made to relate the model to a real data set

this usually takes the form of making a general prediction about the kind of behaviour that will be observed, then seeing if the data exhibits that qualitative structure, as was done in Biro et al. [2006b].

This project began with the goal of creating a truly cross-disciplinary piece of work, to build a model designed to predict a specific type of movement data, constructed from the ground up on solid mathematical principles, but always with a biological aim in mind. Most importantly the model was not to be a predictor of qualitative phenomena, nor a simple application of a 'black box' technique to animal tracking data. Instead it should predict the actual observed data, and in doing so inform us about the structure of that data and the processes that generate it.

This was achieved in Chapters 5 and 6. Chapter 5 constructed a model based on a known biological phenomenon that allowed quantitative prediction of flight paths. A distribution of flight paths was created that could be updated in the light of observations so as to make improved predictions. For the first time this allows for quantitative judgement about the shared information content of different flight paths, and by showing that the shared information increases over time I demonstrated that pigeons learn and follow habitual routes. In Chapter 6 I used this framework to show that paths become more predictable when only subsets of the available data are used to make predictions, and thus showed that habitual routes contain discrete highly predictive regions, which I argued correspond to waypoints and are possibly associated with visual landmarks. Therefore we learn about the structure of the data, that most of the information contained in the path is contained in only a small number of regions and that in other regions previous observations can not aid, and may hinder, our attempts to predict new flight paths. This potentially presents a way to escape the impossibility of using supervised learning to discover landmarks (because of the absence of labelled data of known landmarks), by segmenting the data into salient, informative regions in an unsupervised manner.

These two chapters form the crux of this thesis; the argument is that we can and should predict flight paths, that by doing so using the solid foundations of prob-

ability theory we gain precision in our predictions and clarity in the interpretation of observations, and that doing this allows us to identify structure in the data that potentially corresponds to biological theory. The predictability identified in Chapter 5 reveals the existence of underlying route memories that cause different flight paths to be mutually informative. Discovering that paths are best predicted by relatively small subsets of previously observed data, Chapter 6, is strong evidence for navigation based at least in part on waypoints. This confirms the link between route following and pilotage proposed by Biro et al. [2004]. It also represents an extension of the ideas proposed by Roberts et al. [2004] and Guilford et al. [2004], replacing the use of local predictability as a measure of saliency with a measure of how each region contributes to predicting the complete flight path.

As a corollary to the main argument in Chapters 5 and 6, Chapter 7 showed that the predictive distribution can be effectively used to simulate experimental data. This allows simulated flight paths to be created, based on real observed flight paths, for the first time. Here I used this capability to simulate the possible results of pairing pigeons and releasing them together, but other biological hypotheses could equally well be encoded as an extension to the GP model. The GP model can form a backbone, as the basic unit which describes the distribution of flight paths. Additional layers can be built upon it to describe the result of a hypothesised effect on those paths. The key power of the GP model is its ability to place a distribution over paths themselves and not a derived quantity of those paths. The ability to simulate flight paths on the basis of a biological hypothesis could also be useful in the context of more empirical studies, as a means for constructing a 'null hypothesis'. For example, if one were to propose that observed flight paths are surprisingly coincident with some feature of the landscape, simulated paths could provide a control study, unaffected by the landscape, with which to compare the experimental data.

In assessing the relation of this work to other methods of analysing high precision tracking data, a simple phrase neatly describes the philosophical distinction: *Prediction, not description.* The fundamental model in this thesis is based on few

biological principles — inter-path correlation through habitual route following and intra-path correlation through limited acceleration and the fixed start and end points of the path. However, as seen in Chapter 7, this foundation can be used to construct more elaborate models that incorporate a larger number of biological hypotheses. These hypotheses can be tested in a manner true to the scientific method, since for a given hypothesis the model makes quantitative predictions about the principle observable — the flight path. This is superior, both in principle and practice, to observing an effect (e.g., decreasing inter-path variability) and subsequently explaining it by a *post-hoc* hypothesis. In an example such as habitual-route following, it seems that once the inter-path variability is observed the explanation is clear. However, refer back to Chapter 6; how, without making a predictive model, should we distinguish between the hypothesis that a bird follows a continuous route and the alternative hypothesis that the habitual route is composed of isolated waypoints? To explain data is insufficient; a theory, a hypothesis, a model must make testable predictions.

8.1 Known Issues and Future Directions

8.1.1 Problems Associated with using Time as the Input Variable

The position of a bird in flight is a function of time. Because a bird cannot travel backwards in time, and cannot be in two places at once, this satisfies the necessary properties of a function laid out in Chapter 4. When one looks at a flight path however, it is usually shown in purely spatial co-ordinates, either of longitude and latitude or a derived Cartesian projection. It is in these co-ordinates, with the temporal element removed, that the similarities between paths are most readily seen. Birds may repeatedly visit the same locations in space but at entirely different times. At first sight then it would make more sense to model the flight paths in only the spatial co-ordinates, perhaps with latitude as function of longitude or *vice versa*. Unfortunately this contradicts the necessary properties of a function, since the bird is perfectly

capable of visiting the same longitude multiple times with varying latitudes (or the reverse). It is clear that some external input is required that progresses monotonically from the start of the flight to the end. An alternative to time is the total distance travelled. In practice this is similar to using time, simply because the flight speed of the bird over a single flight is rather constant. Furthermore, using time has the advantage that the hyper-parameters of the model refer more accurately to physical constraints such as limited speed and acceleration.

Fortunately, because flights are spatially similar, flight speed is roughly constant and birds rarely stop or circle during flight when experienced, birds *are* generally in very similar spatial positions at equivalent times on different flights. If this were not so the analysis presented here would either be impossible or hugely more complex. Nevertheless this requirement can cause real difficulties. If a bird's flights are not well aligned *temporally* they may be deemed *unpredictable* despite appearing very similar in spatial co-ordinates. Similarly a well defined waypoint, visited time and time again may not be picked up if there is too much variation in the times of those visits. Fortunately this is a conservative error; some real waypoints may fail to be discovered but we can be confident that discovered waypoints are real.

8.1.2 Possible solutions

In developing the methods described in this thesis I have considered potential ways to either reduce or eliminate the problems associated with the use of time as an input variable. The simplest suggestion is to use the proportion of the total path distance instead of time. In fact this produces no discernible difference in results, since the distance travelled is so closely correlated to the time variable. In aiming to detect waypoints, the outline of an alternative methodology is given below. This system has been tried and found to be overly cumbersome in its current form, but could potentially be improved upon to construct a time-free waypoint detection algorithm. Secondly I discuss, in more general terms, the possibility of extending the framework of Gaussian processes to allow both waypoint detection and path prediction to be

done in a purely spatial context.

Detecting Waypoints

Consider the probability of a flight path, $x(t)$, given the knowledge that it passes through a region of space, A. Denoting that knowledge as: '$\rightarrow A$', we have, using Bayes' rule,

$$p(\, x(t) \mid \rightarrow A, I\,) = \frac{P(\rightarrow A \mid x(t), I\,)\, p(\, x(t) \mid I\,)}{P(\rightarrow A \mid I\,)}. \tag{66}$$

The first element of the numerator on the right hand side is an indicator function, which equals one if the path does pass through the A. The second element is the prior probability of $x(t)$, which can be assigned using a GP model; because the evidence for a set of waypoints is established by the ratio of the conditional probability and the prior probability of the paths the prior probability of $x(t)$ will generally cancel out when assessing this evidence. The denominator represents the prior probability that the path will pass through the region of space before being observed. This can be calculated using the GP model, marginalising over all times at which the passing through could occur. Denoting as $\rightarrow_t A$ the proposition that the path passes through at time t, this reduces to a product integral. If the path never passes through A it must not pass through A at every time step, therefore

$$\begin{aligned} P(\neg \rightarrow A \mid I\,) &= \prod_{t=0}^{1} P(\neg \rightarrow_t A \mid I\,)^{dt} \\ &= \prod_{t=0}^{1} [1 - P(\rightarrow_t A \mid I\,)]^{dt} \\ &= \exp \int_0^1 \ln[1 - P(\rightarrow_t A \mid I\,)] dt \end{aligned} \tag{67}$$

$$\implies P(\rightarrow A \mid I\,) = 1 - \exp \int_0^1 \ln[1 - P(\rightarrow_t A \mid I\,)] dt \tag{68}$$

The quantity required to make this calculation, $P(\rightarrow_t A \mid I\,)$, is given by the marginal distribution of the GP at time index t, integrated over the region A. By integrating over the time component using this product integral we obtain an update probability for the path, conditioned purely on a *spatial* waypoint, the region A. It should be clear

that this could be extended to an arbitrary set of regions, $\rightarrow A_1, \rightarrow A_2, \ldots, \rightarrow A_n$. However, as the set of spatial regions grows the complexity of the calculation of the denominator in Equation (66) undergoes a combinatorial explosion and becomes impractical. In addition, the idea of a waypoint in this model is too restrictive; a waypoint is here defined as a region of space that the bird passes through with probability one. In reality a given path may not always pass through such a precisely defined region of space. On each flight the bird may attend to a previously memorised waypoint with greater or lesser accuracy. The model quickly becomes unwieldy when trying to incorporate the idea that only a subset of all flights will pass through a fixed region, or when trying to make the boundaries of that region less absolute. Another concern is that using this model to select waypoints is vulnerable to the charge of *testing hypotheses suggested by the data*. When deciding which regions of space to consider as waypoints, one must determine regions that the paths all pass through. The model then confirms that the paths are more probable if they are constrained to pass through these regions, since the denominator of Equation (66) is guaranteed to be below one. Deciding whether each of these regions is actually a waypoint therefore requires the selection of a prior distribution of waypoint density — the probability that any given region of space might be a waypoint. The number of waypoints eventually detected depends strongly on the value of this prior.

Extending Gaussian processes

A more elegant solution might involve an extension to the Gaussian process framework to consider the distribution of arbitrary curves in space. Rasmussen and Williams [2006] show that Gaussian processes are closely related to the methods of *regularisation* and *spline interpolation*. Spline interpolation aims to minimise the curvature of an interpolating function, subject to the constraints imposed by the interpolation points — the observed data. For a simple one-dimensional function, $x(t)$, this implies

minimising the functional $J(x)$, where

$$J(x) = \int \left(\frac{d^2x(t)}{dt^2} \right)^2 dt. \tag{69}$$

The curve that minimises this functional turns out to be a cubic spline, a piecewise cubic polynomial with a separate cubic function between each datum. The function that minimises $J(x)$ can be seen as an energy minimising curve, where the energy, E, of the curve is proportional to $J(x)$. This provides a 'best fit' interpolation with minimum curvature. If, instead of minimising the energy, the energy of the curve is assumed to follow a Boltzmann distribution we can assign a probability to any function based on its curvature,

$$p(\, x(t) \mid I \,) \propto \exp(-E(x)) = \exp(-\text{constant} \times J(x)). \tag{70}$$

By assigning probabilities to every possible function that passes through the interpolation points we can obtain not just a best fit interpolation, but also a distribution over functions that gives an estimate of the variance, just like in the GP framework.

Consider now extending this a stage further. Imagine instead of functions, we model flight paths as elastic wires, situated on a board and held down by pins in various locations. These pins might represent waypoints. The wire would sit in a position of minimum energy, where its curvature and extension are minimised, within the constraints imposed by the pins — just like the minimum energy curve for a spline. Imagine now that we have a large ensemble of such wires, fixed in the same spatial locations and able to exchange energy with each other. This is the *canonical ensemble* of statistical physics [Wark, 2001]. Now add energy to the whole system, such that the wires begin to move away from their low-energy configuration, and allow the ensemble to stabilise so that the energy distribution of the wires settles into a thermal distribution, the maximum entropy state. Each wire can now represent a possible path through the 'waypoints', and the probability of any given wire adopting a possible path is given by the Boltzmann distribution, proportional to $\exp(-E(\text{path}))$. But these paths need no longer obey the limitations of a function. They are simply

arbitrary spatial curves, with an energy defined by the elasticity, the extension and the curvature. Such a framework could provide a more general way to look at flight paths in an entirely spatial framework, providing two conditions can be met. Firstly one must be able to calculate the energy of an arbitrary curve, without referring to it in a functional way. This ought to be straightforward; in extremum an actual wire could be forced into the configuration and the energy needed to do so could be measured. Secondly, one must be able to calculate the missing constant of proportionality in Equation (70), the *partition function* Z,

$$Z = \int_{E=0}^{\infty} g(E) \exp(-E(\text{path}))dE, \tag{71}$$

where $g(E)$ is the density of states, the number of different paths with energies between E and $E + dE$. So far the inability to calculate $g(E)$ has prevented further investigation of this possibility.

8.1.3 Limitations in Waypoint Identification

Beyond the problems imposed by the use of time as an input variable, waypoint identification has further limitations, some of which are a necessary consequence of the methodology, others potentially soluble. The method deployed in this thesis to identify waypoints relies on birds consistently flying close to a specified geo-stationary location. This method will never be able to detect the use of landmarks that are not used as waypoints; visual cues used from a distance cannot be identified because the algorithm assumes that waypoints are part of the actual flight path. This limitation is a basic property of the algorithm. A further limitation, and one that could potentially be overcome, is the inability to identify *inconsistently* used waypoints. If a bird visits a particular location on every second or third flight that waypoint may not be identified, since the bird's location on those locations is inconsistent with its location at the same time on other flights. In Equation (56) a summation was used, predicting cyclically over all paths to minimise the effect of outliers. This has the effect of reducing the impact if one path is particularly unpredictable, but also potentially hides relatively

large predictability between some paths. As an example, consider analysing four paths, in which two of the paths use one waypoint, while the other two use a second waypoint. The pairwise similarity of paths that share a waypoint will be high, but the predictability of any one path using the three others will be low, since two of the three use a different waypoint. This issue could potentially be addressed by focusing more on the pairwise predictability of the paths and looking for extreme values, where a few paths are mutually very predictable, but dissimilar to the others.

Appendix A

The Gaussian Distribution

A.1 Gaussian Identities

These Gaussian identities [Rasmussen and Williams, 2006, Osborne and Roberts, 2007] are important for the analytical manipulation of many Gaussian process equations.

The Gaussian distribution is defined by the probability density given by

$$\begin{aligned} p(\,\boldsymbol{x} \mid I\,) &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\equiv \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \end{aligned} \tag{A1}$$

where $\boldsymbol{x}$ is the value of the Gaussian vector random variable, of length N, $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix.

The Gaussian distribution has the property that both conditional and marginal distributions derived from it are also Gaussian. Therefore, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian:

$$p(\,\boldsymbol{x},\, \boldsymbol{y} \mid I\,) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{y}}\end{bmatrix}\right), \tag{A2}$$

then their marginal and conditional distributions are respectively:

$$p(\,\boldsymbol{x} \mid I\,) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{x}}), \tag{A3}$$

$$p(\,\boldsymbol{x} \mid \boldsymbol{y},\, I\,) = \mathcal{N}\big(\boldsymbol{x}; \boldsymbol{\mu} + \boldsymbol{\Sigma}_{\mathbf{xy}}{\boldsymbol{\Sigma}_{\mathbf{y}}}^{-1}(\boldsymbol{y}-\boldsymbol{\nu}), \boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{xy}}{\boldsymbol{\Sigma}_{\mathbf{y}}}^{-1}\,\boldsymbol{\Sigma}_{\mathbf{yx}}\big). \tag{A4}$$

If the mean of a Gaussian distribution is itself Gaussian, then the joint distribution

is also Gaussian. Therefore, if

$$\begin{aligned} p(\,\boldsymbol{x} \mid I\,) &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_x)\,, \\ p(\,\boldsymbol{y} \mid \boldsymbol{x}, I\,) &= \mathcal{N}\big(\boldsymbol{y}; \mathbf{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_{y|x}\big)\,, \end{aligned} \tag{A5}$$

then the joint distribution can be written as

$$p(\,\boldsymbol{x},\, \boldsymbol{y} \mid I\,) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} ; \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\,\boldsymbol{\mu} + \boldsymbol{b} \end{bmatrix} , \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x \mathbf{A}^\mathsf{T} \\ \mathbf{A}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\,\mathbf{A}^\mathsf{T} \end{bmatrix} \right) \tag{A6}$$

and so, using (A3) and (A4)

$$p(\,\boldsymbol{y} \mid I\,) = \mathcal{N}\big(\boldsymbol{y}; \mathbf{A}\,\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\,\mathbf{A}^\mathsf{T}\big) \tag{A7}$$

$$p(\,\boldsymbol{x} \mid \boldsymbol{y},\, I\,) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu} + \boldsymbol{\Gamma}(\boldsymbol{y} - \mathbf{A}\,\boldsymbol{\mu} - \boldsymbol{b}), \boldsymbol{\Sigma}_x - \boldsymbol{\Gamma}\,\mathbf{A}\,\boldsymbol{\Sigma}_x) \tag{A8}$$

where

$$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_x \mathbf{A}^\mathsf{T} \left(\boldsymbol{\Sigma}_{y|x} + \mathbf{A}\,\boldsymbol{\Sigma}_x\,\mathbf{A}^\mathsf{T}\right)^{-1} \tag{A9}$$

Appendix B

Numerical Methods

B.1 Numerical integration via Markov Chain Monte Carlo

Bayesian analysis makes frequent use of integration as a mathematical method to incorporate all the uncertainty in a variable through marginalisation (see Chapter 3). Often the necessary integrals are not solvable in a closed algebraic form. In these cases we must estimate the value of the integral through numerical means.

The task is to estimate the value of F

$$F = \int_V f(\theta) p(\,\theta \mid I\,)\,\mathrm{d}\theta \tag{B1}$$

where θ is typically a (possibly many-dimensional) model parameter and V is the volume of the parameter space we wish to integrate over. The probability density function $p(\,\theta \mid I\,)$ will in general represent either the prior or posterior distribution for the model parameter.

To estimate the value of F we can take *samples* of the function, $f(\theta)$ at sample points (values of θ) of our choosing. Let these sample points be denoted as $\boldsymbol{\theta}$ and the corresponding function evaluations be denoted as $f(\boldsymbol{\theta})$.

B.1.1 Monte Carlo integration

In multi-dimensional input spaces Monte Carlo methods are the preferred means for estimating the integral. The so-called 'curse of dimensionality' expresses the difficulty

in adequately sampling from high dimensional spaces. Since uniform coverage of the parameter space with an acceptable density is not feasible in such spaces we must aim to take samples from the most informative areas of the space. In the case of the integral in equation (B1), if we do not have any belief about the form of $f(\theta)$ (Monte Carlo integration is a non-Bayesian technique), then a convenient choice of sampling distribution is the probability distribution for θ. This will ensure we take most of our samples in the regions where $p(\,\theta \mid I\,)$ is large. It also has the minor benefit of simplifying the form for the estimation equation. If we have samples,$f(\boldsymbol{\theta})$, at sample points drawn from $p(\,\theta \mid I\,)$ we can estimate the value of F by $\hat{F}$,

$$\hat{F} \simeq \frac{1}{N} \sum_{\mathrm{i}=1}^{N} f(\theta_i). \tag{B2}$$

If $p(\,\theta \mid I\,)$ is a pre-defined prior distribution this may be all that is required to estimate the integral. Typically though we are interested in integrals where the probability distribution over θ is a potentially complicated posterior distribution that we cannot draw samples from in a straightforward manner. Importance sampling (see MacKay [2003] Chapter 29) can be used to compensate by reweighting samples drawn from a simpler distribution. Generally a superior alternative, which is used in this thesis, is to apply a Markov Chain Monte Carlo algorithm to draw samples from the desired distribution without reweighting. This avoids the problem of choosing a sampling distribution that approximates the desired distribution and the resulting problems if the approximation is poor, whereby a small subset of the samples comes to dominate the integral.

B.1.2 Markov Chain Monte Carlo (MCMC)

MCMC denotes a class of algorithms that allow samples to be drawn from a chosen probability distribution. An MCMC algorithm specifies a Markov chain (see MacKay [2003]) which, by construction, has the desired sampling distribution as the steady-state solution. The Markov chain is then run until it converges to the steady-state distribution. After convergence the subsequent values of the chain will be samples

from the desired distribution The most widely used is the Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970]. A starting position, θ_0 is chosen, usually at random. The algorithm progresses by proposing an updated value, θ_{prop}, drawn as a sample from a proposal distribution,

$$q(\,\theta_{\text{prop}} \mid \theta_i\,)\,, \tag{B3}$$

where θ_i is the current value. The choice of proposal distribution determines how the parameter space will be explored. In this thesis Gaussian proposal distributions were used to sample log-transformed parameters, equivalent to using heavy-tailed log-Normal distributions to sample the untransformed parameters.

The new value is either accepted or rejected based on the relative probabilities of the current value and new value under both the proposal distribution and the desired sampling distribution. First we calculate the quantity, R

$$R = \frac{p(\,\theta_{\text{prop}} \mid I\,)\,q(\,\theta_i \mid \theta_{\text{prop}}\,)}{p(\,\theta_i \mid I\,)\,q(\,\theta_{\text{prop}} \mid \theta_i\,)} \tag{B4}$$

The proposed value, θ_{prop} is accepted if R is greater than a random number drawn uniformly between 0 and 1 (this implies the proposal is always accepted if $R > 1$),

$$\theta_{i+1} = \theta_{\text{prop}} \quad \text{if} \quad R > r, \quad r \sim U(0,1) \tag{B5}$$

else,

$$\theta_{i+1} = \theta_i. \tag{B6}$$

Convergence to the steady-state distribution is assessed by inspection since there is no standard numerical test for convergence of a Markov chain. A widely used method employed in this thesis is to start several chains at a variety of random starting positions and observe their convergence to a common area of the parameter space.

Bibliography

M. Abramowitz and I. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Dover Publications, 1965.

H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 1974.

C. Armstrong, H. Wilkinson, J. Meade, D. Biro, and T. Guilford. A new role for the sun as a time-compensated landmark in avian navigation. Unpublished.

C. Armstrong, R. Mann, M. Collett, R. Freeman, S. Roberts, H. Wilkinson, and T. Guilford. Why do pigeons form habitual routes? In *Proceedings of the Royal Institute of Navigation Conference*, 2008.

R. Baker. *Migration: paths through time and space*. Hodder and Stoughton, 1982.

R. Baker. *Bird navigation: the solution of a mystery?* Holmes and Meier Publishers, 1984.

N. Baldaccini, S. Benvenuti, V. Fiaschi, P. Ioalè, and F. Papi. Pigeon homing: effects of manipulation of sensory experience at home site. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 94(2):85–96, 1974.

N. Baldaccini, S. Benvenuti, V. Fiaschi, and F. Papi. Pigeon navigation: effects of wind deflection at home cage on homing behaviour. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 99(3):177–186, 1975.

F. Barraquand and S. Benhamou. Animal movements in heterogeneous landscapes: Identifying profitable places and homogeneous movement bouts. *Ecology*, 89(12):3336–3348, 2008.

W. Beck and W. Wiltschko. The magnetic field as a reference system for genetically encoded migratory direction in pied flycatchers (Ficedula hypoleuca Pallas). *Z. Tierpsychol*, 60: 41–46, 1982.

S. Benvenuti and H. Wallraff. Pigeon navigation: site simulation by means of atmospheric odours. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 156(6):737–746, 1985.

S. Benvenuti, V. Fiaschi, L. Fiore, and F. Papi. Homing performances of inexperienced and directionally trained pigeons subjected to olfactory nerve section. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 83(1):81–92, 1973.

G. Bergman and K. O. Donner. An analysis of the spring migration of the common scoter and the long-tailed duck in southern finland. *Acta Zoologica Fennica*, 105:1–59, 1964.

V. P. Bingman and P. Ioalè. Initial orientation of anosmic homing pigeons based on information gathered at familiar release sites remains homeward directed following clock-shift. *Behaviour*, 110:205–218, 1989.

D. Biro. *The role of familiar landmarks in the homing pigeon's familiar area map*. PhD thesis, University of Oxford, 2002.

D. Biro, T. Guilford, G. Dell'Omo, and H. Lipp. How the viewing of familiar landscapes prior to release allows pigeons to home faster: evidence from GPS tracking. *Journal of Experimental Biology*, 205(24):3833–3844, 2002.

D. Biro, J. Meade, and T. Guilford. Familiar route loyalty implies visual pilotage in the homing pigeon. *Proceedings of the National Academy of Sciences of the U.S.A.*, 101(50): 17440–17443, 2004.

D. Biro, J. Meade, and T. Guilford. Route recapitulation and route loyalty in homing pigeons: Pilotage from 25 km? *Journal of Navigation*, 59(01):43–53, 2006a.

D. Biro, D. Sumpter, J. Meade, and T. Guilford. From compromise to leadership in pigeon homing. *Current Biology*, 16(21):2123–2128, 2006b.

D. Biro, R. Freeman, J. Meade, S. Roberts, and T. Guilford. Pigeons combine compass and landmark guidance in familiar route navigation. *Proceedings of the National Academy of Sciences of the U.S.A.*, 104(18):7471–7476, 2007.

F. Bonadonna, L. Dall'Antonia, P. Ioalè, and S. Benvenuti. Pigeon homing: the influence of topographical features in successive releases at the same site. *Behavioural Process*, 39: 137–147, 1997.

V. A. Braithwaite. When does previewing the landscape affect pigeon homing? *Ethology*, 95:141–151, 1993.

V. A. Braithwaite and T. Guilford. Viewing familiar landscapes affects pigeon homing. *Proceedings of the Royal Society B: Biological Sciences*, 245:183–186, 1991.

V. A. Braithwaite and J. A. Newman. Exposure to familiar visual landmarks allows pigeons to home faster. *Animal Behaviour*, 48:1482–1484, 1994.

M. Bramanti, P. Dall'Antonia, and F. Papi. A new technique to monitor the flight paths of birds. *Journal of Experimental Biology*, 134:467–472, 1988.

B. J. Brewer and M. J. Francis. Entropic priors and bayesian model selection. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conf.Proc.*, 2009.

C. Bürgi and S. Werffeli. GPS System zur Aufzeichnung des Flugweges bei Brieftauben. Master's thesis, Institute for Electronics, Swiss Federal Institute of Technology, 1999.

T. Burt, R. Holland, and T. Guilford. Further evidence for visual landmark involvement in the pigeons familiar area map. *Animal Behaviour*, 53:1203–1209, 1997.

J. F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8:679–714, 1986.

I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433:513–516, 2005.

R. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, 1946.

R. T. Cox. *The Algebra of Probable Inference*. Johns Hopkins University Press, Baltimore, MD, 1961.

P. Dall'Antonia, L. Dall'Antonia, and A. Ribolini. Flight path reconstruction of birds by a route recorder. In *Mancini P. et al., Edits. Biotelemetry XII, Proceedings of the XII International Symposium on Biotelemetry. Pisa: Litografia Felici*, pages 544–549, 1993.

T. Dennis, M. Rayner, and M. Walker. Evidence that pigeons orient to geomagnetic intensity during homing. *Proceedings of the Royal Society B: Biological Sciences*, 274(1614):1153, 2007.

P. Fauchald and T. Tveraa. Using first-passage time in the analysis of area-restricted search and habitat selection. *Ecology*, 84(2):282–288, 2003.

R. Freeman. *Analysis of Avian Navigation*. PhD thesis, University of Oxford, 2009.

E. Füller, U. Kowalski, and R. Wiltschko. Orientation of homing pigeons: compass orientation vs piloting by familiar landmarks. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 153(1):55–58, 1983.

A. Gagliardo, P. Ioalè, F. Odetti, and V. P. B. . The ontogeny of the homing pigeon navigational map: evidence for a sensitive learning period. *Proceedings of the Royal Society B: Biological Sciences*, 268:197–202, 2001.

L. Graue. The effect of phase shifts in the day-night cycle on pigeon homing at distances of less than one mile. *Ohio J. Sci*, 63:214–217, 1963.

T. Guilford, S. Roberts, D. Biro, and I. Rezek. Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state models. *Journal of Theoretical Biology*, 227(1):25–38, 2004.

T. Guilford, J. Meade, J. Willis, R. Phillips, D. Boyle, S. Roberts, M. Collett, R. Freeman, and C. Perrins. Migration and stopover in a small pelagic seabird, the Manx shearwater Puffinus puffinus: insights from machine learning. *Proceedings of the Royal Society B: Biological Sciences*, 276(1660):1215–1223, 2009. doi: 10.1098/rspb.2008.1577.

J. Halpern. *Reasoning about uncertainty*. MIT press, 2003.

W. J. Hamilton. *Animal Orientation and Navigation.*, chapter Social aspects of bird orientation mechanisms, pages 57–71. Oregon State University Press, 1967.

R. Hartwick, A. Foa, and F. Papi. The effect of olfactory deprivation by nasal tubes upon homing behavior in pigeons. *Behavioral Ecology and Sociobiology*, 2(1):81–89, 1977.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

C. Haugh, M. Davison, M. Wild, and M. Walker. P-GPS (Pigeon Geomagnetic Positioning System): I. Conditioning analysis of magnetoreception and its mechanism in the homing pigeon (Columba livia). In *Proceedings of the Royal Institute of Navigation Conference*, 2001.

R. Holland. The role of visual landmarks in the avian familiar area map. *Journal of Experimental Biology*, 206(11):1773–1778, 2003.

P. Ioalè, F. Papi, V. Fiaschi, and N. Baldaccini. Pigeon navigation: effects upon homing behaviour by reversing wind direction at the loft. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 128(4):285–295, 1978.

P. Ioalè, M. Nozzolini, and F. Papi. Homing pigeons do extract directional information from olfactory stimuli. *Behavioral Ecology and Sociobiology*, 26(5):301–305, 1990.

P. Ioalè, M. Savini, and A. Gagliardo. Pigeon homing: The navigational map developed in adulthood is based on olfactory information. *Ethology*, 114:95–102, 2008.

E. Jaynes. *Probability Theory*. Cambridge University Press New York, 2003.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 1939.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 186 (1007): 453–461, 1946.

P. E. Jorge, P. A. M. Marques, and J. B. Phillips. Activational effects of odours on avian navigation. *Proceedings of the Royal Society B: Biological Sciences*, 2009.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering*, (82 (Series D)):35–45, 1960.

W. Keeton. Magnets interfere with pigeon homing. 68(1):102–106, 1971.

W. Keeton. Release-site bias as a possible guide to the map component in pigeon homing. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 86(1):1–16, 1973.

W. Keeton. The orientational and navigational basis of homing in birds. *Advances in the Study of Behavior*, 5:47–132, 1974.

J. Kiepenheuer. Pigeon homing: A repetition of the deflector loft experiment. *Behavioral Ecology and Sociobiology*, 3(4):393–395, 1978.

J. Kiepenheuer. The ambiguity of initial orientation of homing pigeons. In *Proceedings of the Royal Institute of Navigation Conference*, 1993.

G. Kramer. Orientierte zugaktivität gekäfigten singvögel. *Naturwissenschaften*, 37:188, 1950.

G. Kramer. Experiments on bird orientation. *Ibis*, 94(265):5, 1952.

G. Kramer. Wird die sonnenhöhe bei der heimfindeorientierung verwertet? *Journal für Ornithologie*, 94:201–219, 1953.

K. Lau, S. Roberts, D. Biro, R. Freeman, J. Meade, and T. Guilford. An edge-detection approach to investigating pigeon navigation. *Journal of Theoretical Biology*, 239(1):71–78, 2006.

N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. *Advances in neural information processing systems*, pages 625–632, 2003.

H. P. Lipp, A. L. Vyssotski, D. P. Wolfer, S. Renaudineau, M. Savini, G. Tröster, and G. Dell'Omo. Pigeon homing along highways and exits. *Current Biology*, 14:1239–1249, 2004.

D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

J. Meade, D. Biro, and T. Guilford. Homing pigeons develop local route stereotypy. *Proceedings of the Royal Society B: Biological Sciences*, 272:17–23, 2005.

J. Meade, D. Biro, and T. Guilford. Route recognition in the homing pigeon, columba livia. *Animal behaviour*, 72:975–980, 2006.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6): 1087–1092, 1953. doi: 10.1063/1.1699114.

M. Osborne and S. Roberts. Gaussian processes for prediction. Technical report, Pattern Analysis and Machine Learning Research Group, University of Oxford, 2007.

F. Papi. Olfactory navigation in birds. *Experientia*, 46:352–363, 1990.

F. Papi, L. Fiore, V. Fiaschi, and S. Benvenuti. The influence of olfactory nerve section on the homing capacity of carrier pigeons. *Monit. zool. ital.(NS)*, 5:265–267, 1971.

F. Papi, L. Fiore, V. Fiaschi, and S. Benvenuti. Olfaction and homing in pigeons. *Monit. zool. ital.(NS)*, 6:85–95, 1972.

F. Papi, G. Mariotti, A. Foa', and V. Fiaschi. Orientation of anosmatic pigeons. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 135(3):227–232, 1980.

T. Patterson, L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. State–space models of individual animal movement. *Trends in Ecology & Evolution*, 2008.

W. Penny, K. Stephan, A. Mechelli, and K. Friston. Comparing dynamic causal models. *NeuroImage*, 22(3):1157–1172, 2004.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008. URL `http://matrixcookbook.com/`.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, 53(3):267–296, 1990.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The M.I.T Press, 2006.

S. Roberts, T. Guilford, I. Rezek, and D. Biro. Positional entropy during pigeon homing. I. application of Bayesian latent state modelling. *Journal of Theoretical Biology*, 227(1): 39–50, 2004.

K. Schmidt-Koenig. *Avian orientation and navigation*. Academic Press London, 1979.

K. Schmidt-Koenig. The sun compass. *Cellular and Molecular Life Sciences (CMLS)*, 46 (4):336–342, 1990.

K. Schmidt-Koenig and H. J. Schlichte. Homing in pigeons with impaired vision. *Proceedings of the National Academy of Sciences of the U.S.A.*, 69 (9):24462447, 1972.

K. Schmidt-Koenig and C. Walcott. Tracks of pigeons homing with frosted lenses. *Animal Behaviour*, 26:480–486, 1978.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464., 1978.

M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, 2003.

A. M. Simons. Many wrongs: the advantage of group navigation. *Trends in Ecology & Evolution*, 19(9):453 – 455, 2004. doi: DOI:10.1016/j.tree.2004.07.001.

M. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

D. Sumpter. *Collective Animal Behaviour*. Published Online, 2008. URL `http://www.collective-behavior.com/`.

D. Sumpter, J. Buhl, D. Biro, and I. Couzin. Information transfer in moving animal groups. *Theory in Biosciences*, 127(2):177–186, 2008.

T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226–1229, 1995.

J. Waldvogel, S. Benvenuti, W. Keeton, and F. Papi. Homing pigeon orientation influenced by deflected winds at home loft. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 128(4):297–301, 1978.

H. Wallraff. Avian olfactory navigation: its empirical foundation and conceptual state. *Animal Behaviour*, 67(2):189–204, 2004.

H. Wallraff and M. Neumann. Contribution of olfactory navigation and non-olfactory pilotage to pigeon homing. *Behavioral Ecology and Sociobiology*, 25(4):293–302, 1989.

H. G. Wallraff. Social interrelations involved in migratory orientation of birds: Possible contribution of field studies. *Oikos*, 30(2):401–404, 1978.

H. G. Wallraff. Initial orientation of homing pigeons as affected by the surrounding landscape. *Ethology Ecology Evolution*, 6:23–26, 1994.

H. G. Wallraff. Seven theses on pigeon homing deduced from empirical findings. *Journal of Experimental Biology*, 199:105–111, 1996.

H. G. Wallraff. *Avian Navigation: Pigeon Homing As A Paradigm*. Springer, 2005.

H. G. Wallraff, J. Chappell, and T. Guilford. The roles of the sun and the landscape in pigeon homing. *Journal of Experimental Biology*, 202:2121–2126, 1999.

J. Wark. *Statistical Mechanics: A Survival Guide*. Oxford University Press, 2001.

M. Wikelski and F. Rienks. Global satellite tracking of (small) animals will revolutionize insight into migration, human health, and conservation. Technical report, ICARUS Initiative, 2008. URL `www.icarusinitiative.org`.

H. Wilkinson, C. Armstrong, J. Meade, and T. Guilford. The role of the sun compass in the familiar area in sight of the loft. In *Proceedings of the Royal Institute of Navigation Conference*, 2008.

R. Wiltschko and W. Wiltschko. The development of sun compass orientation in young homing pigeons. *Behavioral Ecology and Sociobiology*, 9(2):135–141, 1981.

R. Wiltschko and W. Wiltschko. Avian navigation: from historical to modern concepts. *Animal Behaviour*, 65:257–272, 2003.

R. Wiltschko, B. Siegmund, and K. Stapput. Navigational strategies of homing pigeons at familiar sites. *Behavioral Ecology and Sociobiology*, 59:303–312, 2005.

R. Wiltschko, I. Schiffner, and B. Siegmund. Homing flights of pigeons over familiar terrain. *Animal Behaviour*, 74(5):1229–1240, 2007.

W. Wiltschko and E. Gwinner. Evidence for an innate magnetic compass in garden warblers. *Naturwissenschaften*, 61(9):406–406, 1974.

W. Wiltschko and R. Wiltschko. Magnetic Orientation in Birds. *Journal of Experimental Biology*, 199:29–38, 1996.

W. Wiltschko and R. Wiltschko. Magnetic orientation and magnetoreception in birds and other animals. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 191(8):675–693, 2005.

H. Yeagley. A preliminary study of a physical basis of bird navigation. *Journal of Applied Physics*, 18:1035–1063, 1947.

H. Yeagley. A preliminary study of a physical basis of bird navigation II. *Journal of Applied Physics*, 22:746–760, 1951.

M. Zapka, D. Heyers, C. Hein, S. Engels, N. Schneider, J. Hans, S. Weiler, D. Dreyer, D. Kishkinev, J. Wild, et al. Visual but not trigeminal mediation of magnetic compass information in a migratory bird. *Nature*, 461(7268):1274–1277, 2009.