Statistical Methods for Big Tracking Data

by

Yang Liu

B.Sc., Nankai University, 2010

M.Sc., The University of British Columbia, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES

(Statistics)

The University of British Columbia

(Vancouver)

March 2017

Abstract

Recent advances in technology have led to large sets of tracking data, which brings new challenges in statistical modeling and prediction. Built on recent developments in Gaussian process modeling for spatio–temporal data and stochastic differential equations (SDEs), we develop a sequence of new models and corresponding inferential methods to meet these challenges. We first propose Bayesian Melding (BM) and downscaling frameworks to combine observations from different sources. To use BM for big tracking data, we exploit the properties of the processes along with approximations to the likelihood to break a high dimensional problem into a series of lower dimensional problems. To implement the downscaling approach, we apply the integrated nested Laplace approximation (INLA) to fit a linear mixed effect model that connects the two sources of observations. We apply these two approaches in a case study involving the tracking of marine mammals. Both of our frameworks have superior predictive performance compared with traditional approaches in both cross–validation and simulation studies.

We further develop the BM frameworks with stochastic processes that can reflect the time varying features of the tracks. We first develop a conditional heterogeneous Gaussian Process (CHGP) but certain properties of this process make it extremely difficult to perform model selection. We also propose a linear SDE with splines as its coefficients, which we refer to as a generalized Ornstein-Ulhenbeck (GOU) process. The GOU achieves flexible modeling of the tracks in both mean and covariance with a reasonably parsimonious parameterization. Inference and prediction for this process can be computed via the Kalman filter and smoother. BM with the GOU achieves a smaller prediction error and better credibility intervals in cross validation comparisons to the basic BM and downscaling models.

Following the success with the GOU, we further study a special class of SDEs called the potential field (PF) models, which formulates the drift term as the gradient of another function. We apply the PF approach to modeling of tracks of marine mammals as well as basketball players, and demonstrate its potential in learning, visualizing, and interpreting the trends in the paths.

Preface

This dissertation is the original intellectual work of the author Yang Liu under the supervision of Prof. James V. Zidek.

A version of Chapter 2 and Chapter 4 has been published [Liu Y, Battaile BC, Trites AW, Zidek JV, Bias correction and uncertainty characterization of dead-reckoned paths of marine mammals, Animal Biotelmetry 3:51, 2015] and [Liu Y, Zidek ZV, Trites AW, Battaile BC, Bayesian data fusion approaches to predicting spatial tracks: Application to marine mammals, The Annals of Applied Statistics, Vol 10, No.3 1517-1546, 2016]. I was the lead investigator, responsible for all methods development, algorithm design and implementation, simulation design and analysis, data analysis, and manuscript preparation. Battaile BC and Trites AW provided data for the analysis and background of the scientific study. Zidek JV was the supervisory author on this project and was involved throughout the project in concept formation and manuscript edits.

A version of Chapter 7 has also been published [Liu Y, Dinsdale DR, Jun SH, Briercliffe C, Bone J, Statistical learning of basketball strategy via SportVU tracking data: The potential field approach, Cascadia Symposium on Statistics in Sports, 2016]. I was the lead investigator, responsible for most methods development, algorithm design and implementation, data processing and analysis, and manuscript composition. Dinsdale DR was involved in the concept formation. All the other authors were involved in the results visualization, interpretation and manuscript edits.

Table of Contents

List of Tables

List of Figures

Glossary

Some of the frequently used acronyms are listed here.

BB Brownian Bridge

BF Bayes Factor

BM Bayesian Melding

CRAWL Continuous-Time Correlated Random Walk Library

CI Credible Interval

DLM Dynamic Linear Model

DR Dead–Reckoning

DRA Dead–Reckoning Algorithm

GOU Generalized Ornstein–Ulhenbeck

GP Gaussian Process

GPS Global Positioning System

INLA Integrated nested Laplace approximation

IOU Integrated Ornstein–Ulhenbeck

KL Kullback–Leibler

LOOCV Leave–one–out Cross–Validation

L5OCV Leave–five–out Cross–Validation

LR Likelihood Ratio

MLE Maximum Likelihood Estimate

OU Ornstein–Uhlenbeck

RMSE Root Mean Squared Error

SDE Stochastic Differential Equations

Acknowledgments

My deepest debt of gratitude for profound advice and warm support goes to my supervisor, Professor James V. Zidek. As a PhD student, I am blessed to have worked with Prof. Zidek. Through him, I learned many things, such as how to decompose the complex data problem into small problems and tackle them piece by piece, how to convert our wild ideas into solid scientific research, and how to communicate complex statistical methods, etc. All these will benefit me for the rest of my life.

I would also thank my supervisory committee members, Prof. Nancy Heckman and Prof. Matías Salibián-Barrera for their great help and suggestions. Part of this dissertation is a product from the joint research program between the Department of Statistics and Marine Mammal Research Unit, Institute of Ocean and Fisheries at UBC. Prof. Andrew W. Trites and Dr. Brian C. Battaile provided the tracking data of northern fur seals and helped me to understand the scientific background as well as the behavior of those animals. It is great pleasure for me to work with all of them. Through my seven years at UBC for both my master and PhD, I have benefited so much from being immersed in the great research atmosphere in the Department of Statistics. I would thank every member of this department for their company and encouragement through my program. Especially, I would like to thank Prof. Jiahua Chen, Prof. William Welch, and Prof. Harry Joe for many helpful discussions.

My research is partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) via the postgraduate scholarship–Doctoral (PGS-D). I am grateful for their recognition and support for my research.

Last but least, I want to thank my amazing family, especially my wife An-

gela Jiayun Yao and parents Shouhua Liu and Fengrong Zhang, for their love and support throughout all these years.

Dedication

To my wife and parents.

Chapter 1

Introduction

Recent technological advances have made it feasible to track many things, such as foraging trips of endangered animals (Wilson et al., 2007), the movements of basketball players (Miller et al., 2014), and the spread of infectious diseases (Ginsberg et al., 2009; Lazer et al., 2014b). Foraging trips, for example, reflect the feeding areas of animals, and can be used to identify critical habitat. Similarly, the movements of basketball players can be analyzed to identify playing strategies, and evaluate the players (see e.g., Cervone et al., 2014; Liu et al., 2016). It is also essential to track diseases for controlling epidemics and predicting future outbreaks. Tracking helps us to understand the processes that determine movement patterns, and provides a means to forecast and plan for future changes.

The data collection technologies used for tracking involve observations from both direct and indirect sources. Direct observations are usually accurate but sparse in space and time, such as GPS observations of an animal's location in animal tracking, or the number of infected patients from epidemic reports. In contrast, indirect sources are usually inaccurate but data rich, and can help to fill in the gaps in the direct observations, such as inferred locations between GPS observations (i.e., Dead–Reckoned paths), or the search fraction of flu related queries on Google (Ginsberg et al., 2009). Combining different systems of observations in a statistically rigorous manner can ultimately result in more accurate and higher resolution tracks for objects of interest. This is the primary objective for our study.

After high resolution and accurate tracks have been inferred, it is important

to further model and summarize the trends in them, which helps to convert the tracking data into practical knowledge, such as the animal's habitat preference or the basketball players' strategies. Modeling and learning the trends is another key objective of our study. A good statistical model for the tracks can in turn further improve our methods to combine different systems of observations.

The advancement in data collection also results in tracking data in extraordinary high temporal resolution, like the 16Hz animal tracking data and the 25Hz basketball tracking data, both of which will be studied later in this thesis. Such high resolution tracking easily constitutes very big (long) data sets even for a short period. This data rich environment brings remarkable details to the tracking subjects, but also computational challenges in the analysis and modeling. We aim to design our methods such that they can perform well in the prediction and interpretation, but also have good computational efficiency and scalability.

The rest of this thesis documents our efforts and is organized as follows. Chapter 2 reviews the background of mammal marine tracking and the properties of different systems for making observations. It also includes the details of a northern fur seal tracking study, which provides the data set used in much of our work. Our model and methodology development are built on some recent developments in the Gaussian process modeling and stochastic differential equations, which will be reviewed in Chapter 3. We propose and study two frameworks, Bayesian Melding and downscaling to combine different systems of data in Chapter 4. Both perform well relative to competing methods in the seal tracking data. To further improve stochastic models of the tracks, we propose two new stochastic processes, a conditional heterogeneous Gaussian process (CHGP) and a generalized Ornstein–Ulhenbeck (GOU) process in Chapter 5 and 6 respectively. Only the GOU process succeeds in reaching our modeling objectives. To further improve the interpretation of the tracks, we study the potential field approach for track modeling. It is applied to both the seal and basketball tracking data in Chapter 7. The discussion and future work are included in Chapter 8.

In what follows, we describe the mathematical notations in this thesis. Regular characters, such as t, x, η, denote univariate variables or one dimensional quantities, while bold characters, such as $\mathbf{Y}, \mathbf{C}, \boldsymbol{\theta}$, denote multivariate variables, vectors, or matrices. We use $\{\ldots\}$ to denote a process or a collection of variables, $(\ldots)^T$ to

denote a vector formed by its elements, and, $[\ldots]$ to denote a density function. Due to the large amount of notation involved in different contexts explored in this thesis, we have chosen to use the symbols that seem best in the context, i.e., they should be taken as "local" to each chapter or section. So for example, $\mathbf{y}$ in Section 3.1 and Section 3.3 can have different meanings. We will specify their meaning as they are first used.

Chapter 2

Background on Animal Tracking

Marine biologists have been attaching a variety of different electronic tags to marine animals to track their movements, describe their behaviors and characterize their habitat preferences, e.g., Benoit-Bird et al. (2013a). Interactions of tracked animals with other animals or the environment allow for ecological questions regarding population structure and dynamics to be addressed (Block et al., 2011; Benoit-Bird et al., 2013b). Free–ranging animals can serve as sensors to collect environmental (e.g., oceanographic) data, which is difficult or expensive to obtain via conventional means such as ships or satellites (Boehme et al., 2008, 2010; Nordstrom et al., 2013). These environmental data can help to better understand ecosystems and the impact of climate change.

Accurately determining the locations of animals is a fundamental problem in animal tracking. One means of obtaining locations is to have tags carried by animals communicate with satellite systems such as the Global Positioning System (GPS) or the ARGOS satellites. However, as detailed in Section 2.1 and 2.3, the need to communicate with the satellites dramatically limits the resolution of those devices. One means of collecting data on the animal's movements between observations from a satellite tag is to concurrently deploy a "Dead–Reckoning" (DR) tag consisting of an accelerometer, a magnetometer, a time–depth–recorder (TDR) and other supporting components (Wilson et al., 2007; Nordstrom et al., 2013). Such DR tags can sample at infra–second frequencies (e.g., 16Hz) and provide a detailed record of an animal's movements. Data downloaded from the tag can be processed

by a Dead–Reckoning algorithm (DRA) to reconstruct the Dead–Reckoned (DR) path of the animal (Wilson and Wilson, 1988; Johnson and Tyack, 2003; Wilson et al., 2007). The details of DRA are further described in Section 2.2. Section 2.3 provides the information about two data sets from tracking northern fur seals with both GPS and DR devices, which are used throughout this paper.

2.1 Satellite tags

According to Hofmann-Wellenhof et al. (2012), the basic idea of GPS is to first obtain the sights of over four GPS satellites and then calculate the current location by solving a system of equations based on the time and known positions of those satellites. The ARGOS system works in a similar fashion. The necessity of "seeing" the satellites is an impediment in tracking the animals, especially marine animals. These satellite tags cannot work when the animal is diving under the water, which accounts for a large proportion of their foraging trip. Also, the satellite tags have a higher energy consumption rate than the DR tags. To save battery life, the satellite tags are usually programed to be only turned on for a short period (i.e., half a minute) after a long period of hibernation (i.e., 15 minutes) (Battaile et al., 2015). Thus, satellite systems can only provide a sparse and irregularly spaced record of animal locations.

As sparse as those satellites observations are, they accurately record the animal's locations. Extensive experiments (Bryant, 2007; Hazel, 2009) have been carried out to test the precision of the GPS tags by comparing its readings at some known locations to the locations' latitude and longitude. It is found that errors in latitude and longitude are of mean zero and independently normally distributed with similar standard deviation, which depends on the number of satellites observable to the device. When over ten satellites are available, the standard deviation is around 10 meters while it increases to 70 meters when only five satellites are available.

To fill gaps in the observations from satellite tags, many statistical interpolation methods have been developed. McClintock et al. (2014) provide an extensive review of this literature. Most of the approaches developed are state–space based models either in continuous or discrete time and space, such as the continuous time

correlated random walk (CRAWL, Johnson et al., 2008) or robust state space models (Jonsen et al., 2005). Recently, Fleming et al. (2016) proposed Kriging to interpolate the satellite observations and compared the performance of several Gaussian processes with different covariance structures. Hooten and Johnson (2017) developed a continuous-time stochastic integral equation framework, which can be constructed to mimic behavioral characteristics in realistic paths. However, the interpolation from those models depends heavily on the model assumptions, such as the correlated random walk or the Kriging model. Those assumptions may be difficult to justify in practice and lead to unrealistic predictions, as shown in Section 4.5. It is thus compelling to supplement the satellite observations with other sources of data.

2.2 Dead–Reckoning

Instead of trying to record the location directly, the DR tag is designed to collect information related to the animal's movement, such as the animal's velocity, acceleration and body position (orientation), with a high sampling frequency (e.g., 16 times per second or higher). All of this information can be measured locally by the accelerometer, magnetometer, and time–depth–recorder in the DR tag without connecting to the satellites. These measurements are then fed into the Dead–Reckoning algorithm (DRA) for an "estimate" of the animal's location. The DRA is based on the physical rules that produce measurements in the DR tag. The detailed implementation of the DRA may vary in different applications (Wilson and Wilson, 1988; Elkaim et al., 2006; Wilson et al., 2007), but the basic idea is as follows.

1. Obtain the earth's gravity vector $\mathbf{g}$ and magnetic field vector $\mathbf{m}$ at the locations being studied.

2. Correct the systematic bias in the accelerometer and magnetometer that are caused by inadequate calibration (Grewal et al., 2007).

3. Smooth the corrected accelerometer and magnetometer readings with a running mean or a low-pass filter.

4. Find the animal's orientation (often taken to be the tag's orientation, namely the three-dimensional direction of velocity) by solving Wahba's problem (Wahba, 1965), which is to find the rotation matrix $\mathbf{O}$ that minimizes the l_2 distance between the rotated gravity and magnetic forces ($\mathbf{Og}$ and $\mathbf{Om}$) and the corrected and smoothed accelerometer reading $\tilde{\mathbf{a}}$ and magnetometer readings $\tilde{\mathbf{m}}$, namely,

$$\arg\min_{\mathbf{O}} \left\{ ||\tilde{\mathbf{a}} - \mathbf{Og}||^2 + ||\tilde{\mathbf{m}} - \mathbf{Om}||^2 \right\}.$$

5. Obtain the animal's speed (norm of the velocity vector) by one of the following approaches:

   (a) Assume it is a known constant.

   (b) Assume it is measured by a speed meter (a wheel or paddle), which only measures the speed of the animal with respect to the water, but not to the earth.

   (c) From the data recorded by the TDR, calculate the velocity in the depth direction v_z. Given the animal's orientation $\mathbf{o} = \{o_x, o_y, o_z\}$, the three dimensional velocity is then $\mathbf{v} = \mathbf{o}v_z/o_z$.

6. Starting from a known point, integrate the velocity to obtain the animal's trajectory.

The DR path provides remarkably detailed information about an animal's movements, especially fine–scale fluctuations that the GPS cannot capture. However, the DR path can be biased because of poor measurements of swim speeds, systematic as well as random error in the accelerometer and magnetometer sensors, undocumented animal movements caused by ocean currents, confounding between movement and gravitational acceleration, and discretization in the integration of the speed. All of these factors lead to biases and errors in the DR path (Wilson et al., 2007; Liu et al., 2015), which can be significant if not corrected using the relatively accurate GPS observations (by as much as 100km at the end of a seven–day trip in the case study of the next chapter).

2.2.1 Conventional bias correction for DR path

The conventional approach to correct for DR path bias has been to add a linear bias correction term to the DR path, which directly shifts the DR path to the locations indicated by the GPS observations (Wilson et al., 2007). This approach can be summarized as follows: denote the DR path (in one dimension) by $x_1, x_2, \ldots, x_T$ at times $t = 0, 1, 2, \ldots, T$ and the GPS observations at times 0 and T by y_0, y_T respectively; assume without loss of generality, that $x_0 = y_0 = 0$ and that the corrected path $\hat{\eta}_t$ is calculated as,

$$\hat{\eta}_t = x_t + \frac{t}{T}(y_T - x_T), \tag{2.1}$$

which evenly distributes the bias $y_T - x_T$ over the individual time points. The DR path between two GPS observations is shifted directly to the locations indicated by the GPS observations, namely $\hat{\eta}_0 = y_0$ and $\hat{\eta}_T = y_T$. This procedure is repeated for all the sections separated by the GPS observations to correct the whole path.

Unfortunately, this conventional method to correct for the DR path bias is simplistic, and fails to consider the measurement error in the GPS observations. This conventional method also fails to provide a statement about the uncertainty in the corrected path. As a result, the bio–logging community has concern about the validity of the corrected path (Nordstrom et al., 2013) and has generally been reluctant to assign much significance to reconstructed locations. It is these concerns that prompted us to develop the Bayesian Melding and downscaling approaches as competing statistically rigorous methods for track reconstruction that overcome the limitations of the conventional approach. We sought to correct the biased DR paths and quantify the uncertainty in the corrected paths.

2.3 Data sets in our case study

The application in this paper involves tracking data from two lactating northern fur seals captured and tagged on Bogoslof Island (Alaska, USA) as part of the Bering Sea Integrated Research Program (BSIERP) (Benoit-Bird et al., 2013a; Nordstrom et al., 2013). Two tags were glued to the fur of each seal with five minute epoxy: a DR "Daily Diary" tag and a TDR MK 10–F with Fastloc ® GPS technology (both

manufactured by Wildlife Computers). The accelerometer and magnetometer of the DR tag were set to sample 16 times per second (16Hz) while the TDR pressure sensor sampled at 1Hz. The GPS sensor was programmed to make one attempt to connect with the satellite every 15 minutes.

We produced the DR path for two foraging trips made by the two female seals (denoted as "Trip 1" and "Trip 2") using the "TrackReconstruction" R package on the 16Hz data set. This R package was developed based on Wilson and Wilson (1988); Wilson et al. (2007) and described in detail by Battaile (2014). We later sub–sampled the DR path to various frequencies to fit the computational capacity of methods used to get the results. We also projected the GPS observations as longitude and latitude to Easting and Northing in kilometers (km) in a point–wise fashion as per Wilson et al. (2007).

The two foraging trips made by the fur seals in our study were each approximately 1 week in duration. Trip 1 was 7 days and had 274 valid GPS observations, while Trip 2 lasted about 7.5 days and had 130 GPS observations. A large proportion of the GPS locations had time gaps around 15 minutes (Table 2.1)—the designed time gap for the GPS device to record the locations. However, non–negligible proportions of them (13% in Trip 1 and 30% in Trip 2) were greater than 1 hour, and the longest time gap in both trips was longer than 10 hours. The distance between GPS observations can be found in the Figure 4.9. Therefore, the GPS observations were irregularly spaced in time and space, and the duration of the gaps between them were quite large, making it necessary to incorporate the high resolution DR path.

Table 2.1: The sample quantiles in minutes of the time gaps between two consecutive GPS observations in our data.

	Min	10%	25%	50%	75%	90%	Max
Trip 1	14.75	15.00	15.45	18.40	31.68	82.79	953.65
Trip 2	14.75	15.00	15.05	30.00	113.05	130.77	698.47

Chapter 3

Review of Some Fundamental Statistical Methods

In this chapter, we review fundamental statistical methods that serve as building blocks for our methodology development, including Gaussian processes, stochastic differential equations (SDE) and the Kalman filter/smoother. Due to our need for brevity, we review only the knowledge that is used in or related to our following chapters and thus do not give a complete treatment of these topics.

3.1 Gaussian process in spatio–temporal modeling of big data

A Gaussian process (GP) is a stochastic process $\{\eta(\cdot)\}$, which is indexed by t in our cases. It has the property that its realizations at any finite collection of points t have a joint multivariate Gaussian distribution. The index t can be in the time domain $\mathbb{R}^+$, spatial domain $\mathbb{R}^2$, spatio–temporal domain $\mathbb{R}^2 \times \mathbb{R}^+$, or even higher dimensional spaces. A GP is completely specified by its mean function $m(t) = \mathrm{E}\,\eta(t)$ and covariance function $C(t,t') = \mathrm{Cov}\,(\eta(t),\eta(t'))$. That is, for an index set $\mathscr{T} = \{t_1, t_2, \ldots, t_n\}$, the random vector $\boldsymbol{\eta} = (\eta(t_1), \eta(t_2), \ldots, \eta(t_n))^T$ follows a multivariate Gaussian distribution of mean $\mathbf{m} = m(\mathscr{T})$ and covariance

matrix $\mathbf{C} = C(\mathscr{T}, \mathscr{T})$. The density function of $\boldsymbol{\eta}$ is,

$$\pi(\boldsymbol{\eta}; \mathbf{m}, \mathbf{C}) = (2\pi)^{-n/2} |\mathbf{C}|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{\eta} - \mathbf{m})^T \mathbf{C}^{-1} (\boldsymbol{\eta} - \mathbf{m}) \right), \tag{3.1}$$

where we use π to denote a density function. The inverse of the covariance matrix is called the precision matrix, $\mathbf{Q} = \mathbf{C}^{-1}$. An elegant theoretical result is that a zero entry $Q_{i,j}$ in the precision matrix indicates that $\eta(t_i)$ and $\eta(t_j)$ are conditionally independent given the remaining random variables $\boldsymbol{\eta}_{-ij} = \boldsymbol{\eta} \setminus \{\eta(t_i), \eta(t_j)\}$,

$$\pi(\eta(t_i), \eta(t_j) | \boldsymbol{\eta}_{-ij}) = \pi(\eta(t_i) | \boldsymbol{\eta}_{-ij}) \pi(\eta(t_j) | \boldsymbol{\eta}_{-ij}).$$

For $Q_{i,j} = 0$, the conditional independence property also implies the Markovian property,

$$\begin{aligned} \pi(\eta(t_i) | \boldsymbol{\eta}_{-ij}, \eta(t_j)) =& \frac{\pi(\eta(t_i), \eta(t_j) | \boldsymbol{\eta}_{-ij})}{\pi(\eta(t_j) | \boldsymbol{\eta}_{-ij})} = \frac{\pi(\eta(t_i) | \boldsymbol{\eta}_{-ij}) \pi(\eta(t_j) | \boldsymbol{\eta}_{-ij})}{\pi(\eta(t_j) | \boldsymbol{\eta}_{-ij})} \\ =& \pi(\eta(t_i) | \boldsymbol{\eta}_{-ij}). \end{aligned}$$

There are many other important properties of GP and they have been reviewed systematically by Rue and Held (2005).

GP has been ubiquitously used in many statistical studies and applications, such as random effect models (Rue and Held, 2005), spatio–temporal modeling (Le and Zidek, 2006; Cressie and Wikle, 2015; Shaddick and Zidek, 2015) and machine learning for various regression and classification problems (Williams and Rasmussen, 2006), and the references within. However, it is computationally expensive to use GP for big n problems. In this setting, n denotes the number of points or indexes where the GP is observed, i.e., the number of distinct locations where a spatial GP is observed. Evaluating the likelihood or posterior density usually requires calculating the matrix inverse $\mathbf{C}^{-1}$ or solving a system of linear equations $\mathbf{M}\mathbf{x} = \mathbf{f}$ (usually used to calculate $\mathbf{M}^{-1}\mathbf{f}$), where $\mathbf{M}$ is a matrix of dimension n and $\mathbf{f}$ is a vector of length n. An example of such a quantity appears in Equation (4.9). Calculating $\mathbf{M}^{-1}\mathbf{f}$ for a general $\mathbf{M}$ has the time complexity of $O(n^3)$ and storing the matrix $\mathbf{M}$ requires $O(n^2)$ in memory space (Nychka et al., 2015). These complexities can be formidable for big n, e.g., n exceeding $500,000$ as in our case study,

and that prevents us from directly applying the GP for big n data.

Moreover, those large data sets often display some non-stationary features, e.g., different correlations and variances in different regions, which makes the traditional stationary (isotropic) GP inadequate when modeling these data. Therefore, several ingenious statistical methods have been developed to address the non-stationarity and big n issues in the analysis of spatio–temporal and computer experiments.

In Section 3.1.1, we review methods developed to tackle the big n problem for GP. Section 3.1.2 reviews some methods designed to address the non-stationarity issue. The integrated nested Laplace approximation (INLA), which is used in Chapter 4, is a method designed to handle both big n and non-stationarity problems. We walk through the steps of INLA in Section 3.1.3.

3.1.1 Methods to deal with big data

To tackle big n problems, the following four paradigms are commonly invoked: 1, sparse matrix operations; 2, special covariance structures; 3, lower rank models; 4, likelihood approximations. One or more can be used in developing models and methods. We organize the following reviews based on the most fundamental ideas embraced by the paradigm.

Methods based on sparse matrix operations

The main bottleneck of using the GP for big n data is in the storage and computation of the big matrices, but they can be handled more efficiently if the big matrix is sparse, i.e., a large fraction of its entries are zero. Many special data structures and algorithms have been developed to handle the storage and computation of the sparse matrix (see e.g., Bates and Maechler, 2016). The basic idea to reduce the storage cost is to store the non-zero entries and their indexes. The computation for sparse matrix decomposition, inversion, calculating determinant, etc., can also be simplified based on the zero entries (Pissanetzky, 1984).

However, the covariance matrices of most commonly used GPs are not sparse. The idea of covariance tapering (Furrer et al., 2006; Kaufman et al., 2008) is meant to force a covariance/correlation matrix into a sparse matrix. For example, the co-

variance for $\eta(t), \eta(t')$ is assumed to be zero if $|t-t'|$ is larger than a threshold. However, as pointed out by Banerjee et al. (2008) and Datta et al. (2016), tapering completely destroys the long range correlation and often fails to provide a reasonable approximation to the true covariance matrix.

The long range correlation can still be preserved if we work with a sparse precision matrix instead of the covariance matrix. The inverse of a non-diagonal sparse matrix is often not sparse[1], which means that the correlation can still be non-zero even when $|t-t'|$ is large. A GP with a sparse precision matrix is also known as the Gaussian Markov Random Field (GMRF Rue and Held, 2005; Rue et al., 2009). As discussed at the beginning of this section, the "Markov" in the name denotes the conditional independence property of such a GP. The Brownian Motion, Brownian Bridge, autoregressive, Ornstein–Ulhenbeck processes that we later work with can all be viewed as a GMRF. The GMRF is a fundamental idea that used in the methods developed in Rue et al. (2009); Lindgren et al. (2011); Nychka et al. (2015), which will be reviewed later. Besides sparse matrices, the Kalman filter/smoother with an expanded state space can also be used for the inference of GMRF (Särkkä et al., 2013). It is unclear to us how the Kalman filter/smoother implementation compares to the sparse matrix implementation in terms of computation speed.

Methods based on special structures

Gramacy and Lee (2008) proposed a treed Gaussian process (TGP) model that splits the data domain into different sub-regions as in the Bayesian classification and regression tree (Chipman, 1998) and fits a Gaussian process within each sub-region. This approach avoids the big n problem by partitioning the data and also deals with the non-stationarity via the local process in each sub-region. Nonetheless, the GP in each sub-region is fitted separately, which introduces an unnatural discontinuity into the global process. Although the discontinuity can be alleviated by averaging over the tree space, the averaging procedure and training multiple trees increases the computational burden of this method. Park et al. (2011)

[1]This is an empirical statement. For example, the inverse of a tri-diagonal symmetric matrix is not sparse. We can also construct counter examples such that the inverse of a sparse matrix is still sparse. To the best of our knowledge, there are no theorems that state what kind of sparse matrix has a non-sparse inverse.

modified the TGP by imposing a continuity constraint on the boundaries of the sub-regions and obtained the continuous tree via a complex constrained numerical optimization routine.

The tree structure was also considered in the multi-resolution tree structured models in Huang et al. (2002) and Tzeng et al. (2005). When compared to the TGP, the multi-resolution models avoid searching for the tree split points by recursively partitioning the spatial domain of interest into equal sized sub-regions. The space partitioning is done recursively by sub-regions. A spatial autoregressive process (SAP) was proposed in Huang et al. (2002) and Tzeng et al. (2005) for this tree structure, such that the covariance at two leaf locations equals the variance of the sub-region that is the parent of both leaf locations. This covariance structure enables an efficient "change–of–resolution" Kalman filter/smoother algorithm in evaluating the likelihood and posterior, whose time complexity is of order $O(n\log(n))$. But the SAP results in a "blocky" covariance structure and lacks the flexibility needed in modeling complex processes (Tzeng et al., 2005).

Low rank methods

Another way to tackle the big n issue is to restrict the expensive computations, such as the matrix inversion, to a matrix with lower dimension r. This brings us to the low rank method, which can be viewed as approximating a GP by a linear transformation of a Gaussian random vector of dimension r. To illustrate how such a low rank matrix can simplify the computation, we work with an example of $\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\eta}$ is the Gaussian random vector defined in (3.1) and $\boldsymbol{\varepsilon}$ is also multivariate Gaussian with mean zero and covariance matrix $\sigma^2\mathbf{I}$ ($\mathbf{I}$ is a n by n identity matrix) [2]. It is easy to show that $\mathbf{Y}$ has a covariance matrix of $\mathbf{C}+\sigma^2\mathbf{I}$ and its inverse needs to be calculated when evaluating the likelihood or calculating the conditional distribution of $\boldsymbol{\eta}$ given $\mathbf{Y}$, i.e.,

$$\begin{aligned}\mathrm{E}(\boldsymbol{\eta}|\mathbf{Y}) =& \mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I})^{-1}(\mathbf{Y}-\mathbf{m})\\ \mathrm{Var}(\boldsymbol{\eta}|\mathbf{Y}) =& \mathbf{C}-\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I})^{-1}\mathbf{C}^T\end{aligned}$$

[2] This corresponds to the simple kriging model $Y(t) = \eta(t) + \varepsilon(t)$ (see e.g., Cressie and Wikle, 2015), where $\{\varepsilon(\cdot)\}$ is a white noise process that is independent of $\{\eta(\cdot)\}$ and has variance σ^2.

where $\mathbf{m}$ and $\mathbf{C}$ are the mean and covariance matrix of $\boldsymbol{\eta}$ from (3.1). As discussed at the beginning of this section, the storage and inversion of $\mathbf{C}+\sigma^2\mathbf{I}$ is computational expensive.

In the low rank method, the full rank covariance matrix $\mathbf{C}$ is replaced by $\mathbf{KSK}^T$, where $\mathbf{S}$ is a $r \times r$ covariance matrix and $\mathbf{K}$ is a $n \times r$ matrix. Equivalently, it assumes that the Gaussian random vector $\boldsymbol{\eta} = \mathbf{K}\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is another Gaussian random vector of dimension r. To compute the inverse of $\mathbf{KSK}^T + \mathbf{I}$ (without loss of generality, assume $\sigma^2 = 1$), we can use the Sherman–Morrison–Woodbury formula (Henderson and Searle, 1981) to simplify it into

$$(\mathbf{I}+\mathbf{KSK}^T)^{-1} = \mathbf{I}-\mathbf{K}(\mathbf{S}^{-1}+\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T, \tag{3.2}$$

which replaces the inverse of the original $n \times n$ matrix by the inverse of two $r \times r$ matrices ($\mathbf{S}^{-1}$ and $(\mathbf{S}^{-1}+\mathbf{K}^T\mathbf{K})^{-1}$). When $r << n$, the right-hand-side of (3.2) can be much faster to calculate than the left-hand-side. Also, it can take less space to store the two lower dimensional matrices $\mathbf{K}$ and $\mathbf{S}$ than the big matrix $\mathbf{C}$. In the MCMC inference of Bayesian models, the low rank method can also be used to break a high dimensional density into a sequence of lower dimensional ones, which is easier to sample from (see e.g., Datta et al., 2016). On the other hand, we would like to point out that the low rank method cannot preserve the original dependence structure of $\boldsymbol{\eta}$, so that $\mathbf{K}\boldsymbol{\xi}$ should be viewed as an approximation to the original process $\boldsymbol{\eta}$.

Many different approaches have been developed to construct $\mathbf{K}$ and $\mathbf{S}$ or equivalently find a good approximation of $\boldsymbol{\eta}$ based on $\boldsymbol{\xi}$, including the fixed rank kriging (FRK) method in Cressie and Johannesson (2008), Gaussian Predictive Process (GPP) in Banerjee et al. (2008), Nearest-Neighbor Gaussian Process (NNGP) in Datta et al. (2016), and the multi-resolution Gaussian process (MRGP) in Nychka et al. (2015). These approaches are mostly developed for modeling spatial data with the Gaussian process, but can be generalized to GP defined on other spaces. To use the same terminology as the original papers, we will call the index t "location" in this subsection for their review.

In FRK, a Gaussian process $\{\eta(\cdot)\}$ is approximated by

$$\hat{\boldsymbol{\eta}}(t) = \sum_{i=1}^{r} \phi_i(t)\xi_i = \boldsymbol{\phi}(t)^T \boldsymbol{\xi}, \tag{3.3}$$

where $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \phi_2(\cdot), \ldots, \phi_r(\cdot))^T$ is a vector of r basis functions. The vector $\boldsymbol{\xi}$ is usually defined as the random vector formed by a GP evaluated at a set of r grid locations. The basis function $\phi_i(t)$ often decreases with the increase of the distance between location t and the i-th grid location.

In the GPP approach, the lower-dimensional GP $\boldsymbol{\xi}$ is first defined on a knot set $\mathscr{S}$ of size $r < n$, which may be a subset of the observed locations and the GP process is approximated by

$$\tilde{\boldsymbol{\eta}}(t) = C(t, \mathscr{S})C^{-1}(\mathscr{S}, \mathscr{S})\boldsymbol{\xi}, \tag{3.4}$$

where $C(\cdot,\cdot)$ is the covariance matrix. If $\boldsymbol{\phi}(t)^T = C(t, \mathscr{S})C^{-1}(\mathscr{S}, \mathscr{S})$ and the knots and grids are the same, the FRK and GPP approaches yield exactly the same approximation process. For both FRK and GPP, the dimension r decides the goodness of approximation as well as the computational complexity of the estimation and prediction procedure. For large n, r has to be fairly large for $\hat{\eta}$ or $\tilde{\eta}$ to adequately approximate the true process (Datta et al., 2016), which, once, again, leads to computational burden. Also Stein (2014) pointed out these low rank approaches have poor performance under situations where the neighboring observations are strongly correlated.

To further improve the FRK approach, Nychka et al. (2015) considered a multi-resolution Gaussian process (MRGP), such that

$$\hat{\boldsymbol{\eta}}(t) = \sum_{l=1}^{L} \boldsymbol{\phi}_l(t)^T \boldsymbol{\xi}_l = \sum_{l=1}^{L} \sum_{j=1}^{r_l} \boldsymbol{\phi}_{l,j}(t)^T \xi_{l,j}, \tag{3.5}$$

where $\boldsymbol{\xi}_l = \{\xi_{l,1}, \xi_{l,2}, \ldots, \xi_{l,r_l}\}$ is a Gaussian Markov process defined on a regular grid with spacing δ_l and each $\phi_{l,j}(\cdot)$ is a basis function similar to that in (3.3). This MRGP is a summation of FRK models over different resolutions, which are designed to capture correlations at different scales. In particular, they assume that

$\delta_l = \delta_{l-1}/2$ and the $\boldsymbol{\xi}_l$ at each resolution are mutually independent. With the Markov process, a sparse precision matrix is introduced for the lower dimensional process and reduces the original order of $O(r^3)$ in computational complexity to at most $O(r^2)$, which means that we can afford a larger r than FRK with the same computational power.

In parallel with the MRGP, Datta et al. (2016) modified the GPP into a NNGP model to overcome the aforementioned shortcomings of the low rank approaches. To avoid the computational burden of large r in the knot process, the joint density of $\boldsymbol{\xi}$ on the knot set $\mathscr{S}$ in GPP is approximated by a product of $m-$nearest neighbor conditional distributions, e.g., $\pi(\xi_1,\xi_2,\xi_3) \approx \pi(\xi_1)\pi(\xi_2|\xi_1)\pi(\xi_3|\xi_2)$ for $m = 1$. The nearest neighbor structure helps to reduce the computational complexity in evaluating this density from order $O(r^3)$ to $O(rm^3)$, which is a huge improvement when $m \ll r$. Also, in the MCMC of Bayesian inference for the NNGP, the non-knot locations are updated by conditioning on their m-nearest neighbors in the knot set, such that $C(t,\mathscr{S})C^{-1}(\mathscr{S},\mathscr{S})$ in (3.4) is replaced by $C(t,\mathbf{w}(t))C^{-1}(\mathbf{w}(t)),\mathbf{w}(t))$, where $\mathbf{w}(t)$ are m-nearest neighbors of t in the knot set $\mathscr{S}$. This further reduces the computational burden of the NNGP.

Because the nearest neighbor structure in NNGP can be viewed as a high order Markov structure, a NNGP is equivalent to a MRGP with $L = 1$ and a certain knot set and parameterization of the covariance matrix. However, both the MRGP and NNGP only consider constant variance and correlation parameters, i.e. parameters in the basis functions $\boldsymbol{\phi}_l(\cdot)$ of the MRGP or the covariance function $C(\cdot,\cdot)$ of the NNGP, over the whole study space. This essentially assumes that the same dependence structures are repeatedly observed over the whole space. Therefore, neither the NNGP nor the MRGP has the flexibility to deal with the inhomogeneous covariance structure over different sub-regions of the space. We call a process "inhomogeneous" (heterogeneous) when it has a non-stationary covariance function with non-constant variance/correlation parameters. For example, a standard Brownian Motion is non-stationary but still homogeneous.

Methods based on approximations

The big n problem can also be tackled by constructing approximations to the full likelihood via a sequence of conditional likelihoods as in Vecchia (1988) and Stein et al. (2004). Caragea and Smith (2007) reviewed and compared these approximation methods based on conditional likelihoods. Eidsvik and Shaby (2014) proposed another approximation method based on the pairwise composite likelihood. These likelihood approximation methods are mostly used for inference in adapting a GP model without a sparse precision matrix, such as the Matérn process. An oversimplified view is that they essentially break a large data set into several small ones and only evaluate the likelihood in these small data sets, which omits certain long range correlations. A concern about them is that the conditional distributions used may not result in a proper joint distribution (Datta et al., 2016).

In addition to the conditional/composition likelihood approximation, Lindgren et al. (2011) developed another approximation to the GP with Matérn covariance functions by representing the process as a solution to a Stochastic Partial Differential Equation (SPDE) and then solving for the SPDE with a finite element method. This can be viewed as approximating a non-Markov GP by a GMRF, whose inference can then be handle by INLA. To illustrate the approximation process, let's consider a Matérn process observed on a regular two dimensional n by n grid $t_{1,1} = (1,1), t_{1,2} = (1,2), \ldots, t_{1,n} = (1,n), t_{2,1} = (2,1), t_{2,2} = (2,2), \ldots, t_{n,n} = (n,n)$. For expository simplicity, denote $\eta(t_{i,j})$ as $\eta_{i,j}$. When the smoothness parameter of the Matérn process ν approaches 0 and the grid size n is large enough (such that we can ignore the boundary effects), the Matérn process approximately has the following conditional distribution

$$\mathrm{E}\left(\eta_{i,j}|\boldsymbol{\eta}_{-i,-j}\right) = \frac{1}{a}\left(\eta_{i-1,j} + \eta_{i+1,j} + \eta_{i,j-1} + \eta_{i,j+1}\right), \tag{3.6}$$

$$\mathrm{Var}\left(\eta_{i,j}|\boldsymbol{\eta}_{-i,-j}\right) = \frac{1}{a}. \tag{3.7}$$

where $a > 4$ is a parameter that depends on the range parameter of the Matérn process. Notice that (3.6) and (3.7) define a GMRF through the conditional distribution (Rue and Held, 2005), i.e., $\eta_{i,j}$ is conditionally independent from all the other $\eta_{i',j'}, |i'-i| > 1,$ or $|j'-j| > 1$ given its nearest neighbors. In addition to the

limiting case of $\nu \to 0$, Lindgren et al. (2011) shows that the Matérn process with $\nu = 1$ can be approximated by a GMRF whose conditional distribution is

$$\begin{aligned}
\mathrm{E}\left(\eta_{i,j}|\boldsymbol{\eta}_{-i,-j}\right) =& \frac{2a}{4+a^2}\left(\eta_{i-1,j}+\eta_{i+1,j}+\eta_{i,j-1}+\eta_{i,j+1}\right) \\
&- \frac{1}{4+a^2}\left(\eta_{i-2,j}+\eta_{i+2,j}+\eta_{i,j-2}+\eta_{i,j+2}\right) \\
&- \frac{2}{4+a^2}\left(\eta_{i-1,j-1}+\eta_{i-1,j+1}+\eta_{i+1,j-1}+\eta_{i+1,j+1}\right), \\
\mathrm{Var}\left(\eta_{i,j}|\boldsymbol{\eta}_{-i,-j}\right) =& \frac{1}{4+a^2}.
\end{aligned}$$

Similar approximations can be constructed for other integer νs. For the Matérn process observed at irregular locations, Lindgren et al. (2011) first uses a finite element method to construct a mesh (i.e., divide the space into triangles) and then builds the approximation GMRF on this mesh. The combination of the SPDE and INLA approximation enables efficient computation of GP's for some big spatio–temporal data. It can also offer certain types of non-stationary modeling. However, Nychka et al. (2015) expressed concern about its computational burden for very large data sets and its lack of flexibility in non-stationary modeling.

3.1.2 Methods for non-stationarity

The non-stationarity issue of GP in spatial analysis was first addressed in Sampson and Guttorp (1992). Their idea was to find a spatial deformation via multi-dimensional scaling that maps the non-stationary field into a stationary one. This method was later extended in different ways as in Damian et al. (2001) as well as Schmidt and O'Hagan (2003). Bornn et al. (2012) addressed the non-stationary issue by embedding the geographical plane in some higher-dimensional spaces to restore stationarity. This dimension expansion approach enjoys better flexibility and interpretability compared to the spatial deformation approach. But both approaches essentially add one or more dimensions to the domain over which the data are collected. This intensifies the computational burden in the analysis.

Another approach to non-stationary modeling is the process convolution approach pioneered by Higdon et al. (1999) and further developed in Paciorek and Schervish (2006), Lemos and Sansó (2012) and Chu et al. (2014), etc. This ap-

proach is inspired by the definition of a spatial moving average process $\eta(s)$:

$$\eta(s) = \int k(s-u)x(u)du,$$

where s, u are the locations in a two-dimensional space, $\{x(\cdot)\}$ is a white noise process and $k(\cdot)$ is a convolution kernel that does not depend on location s. Higdon et al. (1999) achieved a non-stationary spatial model by allowing $k(s-u)$ to vary with s as $k_s(u)$, a so–called spatially evolving kernel. Certain parameterizations of the spatially evolving kernel result in a covariance matrix that appears to be a weighted average of two covariance matrices as in Paciorek and Schervish (2006) and Ba and Joseph (2012), where one covariance matrix models the long range dependence and the other captures the short range dependence. The idea of spatially evolving kernel is also used in Hooten and Johnson (2017).

We observe that the non-stationarity issue is usually caused by the parsimonious parameterizations of the model and can be alleviated by allowing more flexible parameterizations of the GP. The problem of non-stationarity then becomes that of finding a good balance between modeling flexibility, identifiability and computational cost.

3.1.3 Basics of the integrated nested Laplace approximation

The integrated nested Laplace approximation (INLA, Rue et al., 2009; Lindgren et al., 2011) is a novel approximation technique in Bayesian inference, especially for latent GMRF models. Basically, the INLA method seeks the posterior mode via numerical optimization, approximates the integrals of the random effects or hyper-parameters via the Laplace approximation, and numerically integrates over the hyper-parameters on a selected grid based on the (approximated) likelihood. When combined with the approximations to SPDE, INLA can handle the inference for most commonly used covariance models in spatio–temporal modeling. In this subsection, we provide a brief review of its basic ideas and readers can refer to Rue et al. (2009), Martins et al. (2013), and Blangiardo and Cameletti (2015) for detailed discussion and examples.

Consider the following latent Gaussian model: the observations $\mathbf{y}$ depend on a latent random process or field $\boldsymbol{\eta}$ through an observation model (it is also called the

likelihood model in Rue et al. (2009)); The latent process is usually assumed to be Gaussian, i.e., $\boldsymbol{\eta}|\boldsymbol{\psi} \sim \mathrm{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\psi}))$, where the covariance matrix $\mathbf{C}$ depends on the hyper-parameters $\boldsymbol{\psi} = \{\psi_1, \psi_2, \ldots, \}^T$. For example,

$$\begin{aligned} \text{Observation Model: } & \mathbf{y}|\boldsymbol{\eta} \sim \mathrm{N}(\boldsymbol{\eta}, \mathbf{C}_y(\psi_2)) \\ \text{Latent Model: } & \boldsymbol{\eta}|\boldsymbol{\psi} \sim \mathrm{N}(\mathbf{0}, \mathbf{C}_\psi(\psi_1)) \\ \text{Priors: } & \pi(\log(\psi_j)) \propto 1, \qquad\qquad j = 1, 2. \end{aligned}$$

Notice that we use $\pi()$ to denote the density functions, including the prior, conditional, or joint densities. With this model, we are usually interested in calculating the posterior density $\pi(\boldsymbol{\psi}|\mathbf{y})$ and the posterior predictive density $\pi(\mathbf{y}^*|\mathbf{y})$. This can be done by Monte Carlo methods, such as MCMC or Gibbs sampler, but they are usually computationally expensive. The INLA approach is developed as a computationally simpler alternative to the Monte Carlo methods.

Nested Laplace approximation

The first building block of INLA is the Laplace approximation, which approximates an arbitrary density function π at its mode $\hat{x}$ via a second order Taylor series approximation,

$$\begin{aligned} \log(\pi(x)) \approx & \log(\pi(\hat{x})) + \frac{\partial \log(\pi(\hat{x}))}{\partial x}(x-\hat{x}) + \frac{1}{2}\frac{\partial^2 \log(\pi(\hat{x}))}{\partial x^2}(x-\hat{x})^2 \\ = & \log(\pi(\hat{x})) + \frac{1}{2}\frac{\partial^2 \log(\pi(\hat{x}))}{\partial x^2}(x-\hat{x})^2. \end{aligned}$$

The second line follows as $\partial \log(\pi(x))/\partial x = 0$ at the mode $\hat{x}$.

Let $\hat{\sigma}^2 = -1/(\partial^2 \log(\pi(\hat{x}))/\partial x^2)$

$$\log(\pi(x)) \approx \log(\pi(\hat{x})) - \frac{1}{2\hat{\sigma}^2}(x-\hat{x})^2.$$

Taking the exponential of the above equation, we find that $\tilde{\pi}$, the density of $\mathrm{N}(\hat{x}, \hat{\sigma}^2)$, can be a good approximation to π, at least in a small neighborhood near $\hat{x}$. The density $\tilde{\pi}$ is referred to as the Laplace approximation density.

To apply this Laplace approximation to the posterior densities of interest, we

first work with the following identity of the conditional probabilities,

$$\pi(\boldsymbol{\eta}|\boldsymbol{\psi},\mathbf{y}) = \frac{\pi(\boldsymbol{\psi},\boldsymbol{\eta}|\mathbf{y})}{\pi(\boldsymbol{\psi}|\mathbf{y})} \Rightarrow \pi(\boldsymbol{\psi}|\mathbf{y}) = \frac{\pi(\boldsymbol{\psi},\boldsymbol{\eta}|\mathbf{y})}{\pi(\boldsymbol{\eta}|\boldsymbol{\psi},\mathbf{y})}. \tag{3.8}$$

The numerator above may not be evaluated in closed form, but it is proportional to the product of three densities in closed form,

$$\pi(\boldsymbol{\psi},\boldsymbol{\eta}|\mathbf{y}) = \frac{\pi(\mathbf{y},\boldsymbol{\psi},\boldsymbol{\eta})}{\pi(\mathbf{y})} \propto \pi(\mathbf{y},\boldsymbol{\psi},\boldsymbol{\eta}) = \pi(\mathbf{y}|\boldsymbol{\psi},\boldsymbol{\eta})\pi(\boldsymbol{\eta}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})$$

The factor $\pi(\mathbf{y})$ does not depend on any of the unknown parameters $\boldsymbol{\eta}$ or $\boldsymbol{\psi}$.

The denominator in (3.8) can be approximated by the Laplace approximation density $\pi(\boldsymbol{\eta}|\boldsymbol{\psi},\mathbf{y}) \approx \tilde{\pi}(\boldsymbol{\eta}|\boldsymbol{\psi},\mathbf{y})$, yielding

$$\pi(\boldsymbol{\psi}|\mathbf{y}) \approx \tilde{\pi}(\boldsymbol{\psi}|\mathbf{y}) = \left.\frac{\pi(\boldsymbol{\psi},\boldsymbol{\eta}|\mathbf{y})}{\tilde{\pi}(\boldsymbol{\eta}|\boldsymbol{\psi},\mathbf{y})}\right|_{\boldsymbol{\eta}=\arg\max \pi(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi})}.$$

When the likelihood model is also Gaussian, $\tilde{\pi}(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi}) = \pi(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi})$ exactly. For other models, such as logistic, Poisson, etc., the Laplace approximation density can be derived by a Taylor expansion of the likelihood function. This gives the "nested" part of the approximation, as a Laplace approximation density $\tilde{\pi}(\boldsymbol{\eta}|\boldsymbol{\psi},\mathbf{y})$ is used to derive another approximate density $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})$.

Another "nested Laplace" approximation in INLA can be used to improve the approximation to $\pi(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi})$ when the likelihood model is not Gaussian. It calculates the Laplace approximation iterating through all the elements of $\boldsymbol{\eta}$

$$\begin{aligned}
\pi(\eta_j|\boldsymbol{\psi},\mathbf{y}) =& \frac{\pi(\eta_j,\boldsymbol{\eta}_{-j}|\boldsymbol{\psi},\mathbf{y})}{\pi(\boldsymbol{\eta}_{-j}|\eta_j,\boldsymbol{\psi},\mathbf{y})} = \frac{\pi(\eta_j,\boldsymbol{\eta}_{-j},\boldsymbol{\psi}|\mathbf{y})}{\pi(\boldsymbol{\psi}|\mathbf{y})}\frac{1}{\pi(\boldsymbol{\eta}_{-j}|\eta_j,\boldsymbol{\psi},\mathbf{y})} \\
&\propto \frac{\pi(\boldsymbol{\eta},\boldsymbol{\psi}|\mathbf{y})}{\pi(\boldsymbol{\eta}_{-j}|\eta_j,\boldsymbol{\psi},\mathbf{y})} \propto \frac{\pi(\mathbf{y}|\boldsymbol{\psi},\boldsymbol{\eta})\pi(\boldsymbol{\eta}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\pi(\boldsymbol{\eta}_{-j}|\eta_j,\boldsymbol{\psi},\mathbf{y})} \\
&\approx \left.\frac{\pi(\mathbf{y}|\boldsymbol{\psi},\boldsymbol{\eta})\pi(\boldsymbol{\eta}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\tilde{\pi}(\boldsymbol{\eta}_{-j}|\eta_j,\boldsymbol{\psi},\mathbf{y})}\right|_{\boldsymbol{\eta}_{-j}=\hat{\boldsymbol{\eta}}_{-j}(\eta_j,\boldsymbol{\psi})} \triangleq \tilde{\pi}(\eta_j|\boldsymbol{\psi},\mathbf{y}),
\end{aligned} \tag{3.9}$$

where $\triangleq$ denotes "define as" and $\hat{\boldsymbol{\eta}}_{-j}(\eta_j,\boldsymbol{\psi}) = \arg\max \tilde{\pi}(\boldsymbol{\eta}_{-j}|\eta_j,\mathbf{y},\boldsymbol{\psi})$. This improved approximation comes at a high computational cost, because it needs to iter-

ate through all elements of $\boldsymbol{\eta}$.

Integration step of INLA

The approximate Laplace densities are put together for the posterior density of $\pi(\boldsymbol{\eta}|\mathbf{y})$.

$$\begin{aligned}\pi(\boldsymbol{\eta}|\mathbf{y}) &= \int_{\boldsymbol{\psi}} \pi(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi} \\ &\approx \int_{\boldsymbol{\psi}} \tilde{\pi}(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi})\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi}.\end{aligned}$$

The integration is done numerically by first searching the surface of $\tilde{\pi}(\psi|\mathbf{y})$ to setup a grid of integration points, i.e., $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_L$. Each grid point is associated with its own weight $w_i \propto \tilde{\pi}(\boldsymbol{\psi}_i|\mathbf{y})\Delta_i, i = 1,2,\ldots,L$, where Δ_i is area size.

$$\pi(\boldsymbol{\eta}|\mathbf{y}) \approx \sum_{i=1}^{L} w_i \tilde{\pi}(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi}_i).$$

This integration scheme is also used in our Bayesian Melding approach in Section 4.2. If we are interested in a better approximation to the posterior marginal of η_j, $\tilde{\pi}(\boldsymbol{\eta}|\mathbf{y},\boldsymbol{\psi})$ can be replaced by $\tilde{\pi}(\eta_j|\boldsymbol{\psi},\mathbf{y})$ from Equation (3.9).

3.2 Stochastic differential equations

Stochastic differential equations have been used in many different areas, such as physics, finance, and ecology (Iacus, 2009). They model the changes in a process through a stochastic model. An SDE is usually written in the form of

$$d\theta(t) = \mu(\theta(t),t)dt + \sigma(\theta(t),t)dW(t), \tag{3.10}$$

where $\{W(\cdot)\}$ is the standard Wiener process, $\mu(\cdot)$ is the deterministic part, $\sigma(\cdot)$ scales the random noise $dW(t)$, and $dW(t)$ is a white noise process. Equation (3.10) should be understood to be an abbreviation for the following integral,

$$\theta(t) = \theta(0) + \int_0^t \mu(\theta(s),s)ds + \int_0^t \sigma(\theta(s),s)dW(s).$$

The third term $\int_0^t \sigma(\theta(s),s)dW(s)$ is a stochastic integral because of the $dW(s)$ term. Recall that the regular Riemann integral is defined as the limit of the summation of step functions with respect to a partition of an interval $[0,t]$. Let $\Pi_n = \{0 = t_{n,0} < t_{n,1} < t_{n,2} < \ldots < t_{n,n} = t\}$ be any partition of the interval $[0,t]$ and $||\Pi_n|| = \max_{i=0,1,\ldots,n-1}(t_{n,i+1} - t_{n,i})$. The stochastic integral is most commonly defined as

$$\int_{t_0}^{t} X(s)dW_s = \lim_{||\Pi_n||\to 0} \sum_{i=0}^{n-1} X(t_{n,i})(W(t_{n,i+1}) - W(t_{n,i})), \tag{3.11}$$

where we replace $\sigma(\theta(s),s)$ with $X(s)$ for notational simplicity.

The limit in (3.11) is the limit in quadratic mean (mean square). This stochastic integral is known as the Itô integral. It is important to notice that $X(\cdot)$ in the interval $[t_i, t_{i+1}]$ is approximated by a step function evaluated at the start of the interval $X(t_i)$ in (3.11). This gives Itô integral properties different from a regular Riemann integral. For example, $\int_0^t W(s)dW(s) = (W(t)^2 - t)/2$, different from the Riemann integral $\int_0^t s\,ds = t^2/2$. Replacing $X(t_i)$ by $X((t_i + t_{i+1})/2)$ leads to a different definition of the stochastic integral. A more formal treatment of the definitions of stochastic integral can be found in Arnold (1974). For the SDE formulated in (3.10), the functions $\mu(\theta,t)$ and $\sigma(\theta,t)$ have to satisfy a few conditions to make sure there exists an unique continuous solution to (3.10). Notice that we use θ to represent the first argument in these two functions. First, $\mu(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ must be measurable on the domain of θ and t, namely $\mathbb{R}^d \times [0,T]$, where d is the dimension of θ. Second, they need to satisfy the global Lipschitz condition such that there exist a constant $K < +\infty$ with

$$|\mu(\theta,t) - \mu(\theta',t)| + |\sigma(\theta,t) - \sigma(\theta',t)| < K|\theta - \theta'| \tag{3.12}$$

for all $\theta, \theta' \in \mathbb{R}^d$ and $t \in [0,T]$. The final condition requires that there exists a constant $C < +\infty$, such that

$$|\mu(\theta,t)| + |\sigma(\theta,t)| < C(|\theta| + 1) \tag{3.13}$$

for all $\theta, \theta' \in \mathbb{R}^d$ and $t \in [0,T]$. This condition ensures that the solution (3.10) does

not explode in a finite time (see e.g., Iacus, 2009). As seen in Arnold (1974), these three conditions also ensure that the solution is a Markovian diffusion process, in which case, $\mu(\theta,t)$ and $\sigma(\theta,t)$ are refereed to as the drift and diffusion coefficients respectively. Heuristically, the Markovian property can be deduced from the fact that the change in the process only depends on the current state, i.e., not on the previous state. The Markovian property guarantees an efficient computation in the inference and prediction of this process.

Modeling the derivative $d\theta(t)$ can conveniently incorporate our knowledge about how the process changes and evolves, but this parameterization only results in closed form solutions in very few cases (Arnold, 1974; Iacus, 2009). One type of SDE with a closed form solution is the linear SDE discussed in Chapter 6. For the general SDE without a closed form solution in Chapter 7, approximation techniques are needed for the inference and prediction. The simplest and commonly used approximation method is the Euler approximation (see e.g., Iacus, 2009). For a general SDE in (3.10), the Euler approximation defines the approximation process $\tilde{\theta}(t_0), \tilde{\theta}(t_1), \ldots, \tilde{\theta}(t_n)$ as

$$\tilde{\theta}(t_{i+1}) - \tilde{\theta}(t_i) = \mu(\tilde{\theta}(t_i), t_i)(t_{i+1} - t_i) + \sigma(\tilde{\theta}(t_i), t_i)(W(t_{i+1}) - W(t_i)),$$

for $i = 0, 1, 2, \ldots, n-1$. The approximation process $\tilde{\theta}$ is a non-standard Brownian Motion process that is locally constant, whose mean and standard deviation are determined by μ and σ. To improve upon the Euler approximation, Ozaki (1992) developed a local linear approximation method, which can be viewed as using a linear SDE to approximate a general SDE. Archambeau et al. (2007), Archambeau and Opper (2010) and Vrettas et al. (2015) proposed a variational algorithm to minimize the Kullback–Leibler (KL) divergence between the true distribution and the approximate distribution. The approximate distribution can be constructed as a GP (Archambeau et al., 2007; Archambeau and Opper, 2010) or a linear SDE (Vrettas et al., 2015). However, we have not found a complete comparison of these approximation techniques in terms of their theoretical/empirical performance. Also, we notice that the most sophisticated approximation technique only approximates the SDE with a linear SDE.

3.3 Kalman filter and smoother

As shown later in Chapter 6, the linear SDE can be expressed as a linear State Space Model (SSM), alternatively known as the Dynamic Linear Model (DLM, see e.g., Petris et al., 2009). In general, it is formulated as

$$\mathbf{Y}_t = \mathbf{F}_t \boldsymbol{\theta}_t + \mathbf{v}_t \qquad \mathbf{v}_t \sim \mathrm{N}_m(\mathbf{0}, \mathbf{V}_t) \tag{3.14}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \mathbf{u}_t \qquad \mathbf{u}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{U}_t), \tag{3.15}$$

when $t > 0$. The initial condition is $\boldsymbol{\theta}_0 \sim \mathrm{N}(\mathbf{m}_0, \mathbf{C}_0)$.

$\mathbf{Y}_t$ is the $q-$dimensional observation vector and $\boldsymbol{\theta}_t$ is the $p-$dimensional state vector. In general, the dimension p of the state vector can be allowed to be different from the dimension q of the observation vector. Equation (3.14) is referred to as the observation model, where $\mathbf{Y}_t$ is a linear transformation of $\boldsymbol{\theta}_t$ by a $q \times p$ matrix $\mathbf{F}_t$ plus a Gaussian noise $\mathbf{v}_t$ with mean zero and covariance $\mathbf{V}_t$. Equation (3.15) describes the state model, where the current state $\boldsymbol{\theta}_t$ is modeled as a linear transformation of the previous state $\boldsymbol{\theta}_{t-1}$ by a $q \times q$ matrix $\mathbf{G}_t$ plus the Gaussian noise $\mathbf{u}_t$ with mean zero and covariance $\mathbf{U}_t$. The state $\boldsymbol{\theta}_{t-1}$ and two Gaussian noises $\mathbf{v}_t$, $\mathbf{u}_t$ are assumed to mutually independent. The initial condition is considered to be fixed i.e., $\mathbf{C}_0 = \mathbf{0}$, in our study.

Notice that the subscripts t for all the notation above, $\mathbf{Y}_t, \boldsymbol{\theta}_t, \mathbf{F}_t, \ldots$ are simplifications of $\mathbf{Y}(t), \boldsymbol{\theta}(t), \mathbf{F}(t), \ldots$ and they can have arbitrary gaps either equal or unequal. We use $\mathbf{y}_t$ to denote the observation of variable $\mathbf{Y}_t$ from the DLM in (3.14) and (3.15) and $\boldsymbol{\psi}$ to denote all the parameters in $\mathbf{m}_0, \mathbf{C}_0, \mathbf{F}_{1:T}, \mathbf{G}_{1:T}, \mathbf{V}_{1:T}, \mathbf{U}_{1:T}$. Notation $a:b$ represents collection of $\{a, a+1, a+2, \ldots, b\}$, i.e., $\boldsymbol{\theta}_{1:t} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_t\}$. For expository simplicity, we use $[\cdots]$ instead of $\pi()$ to denote the density functions hereafter in this thesis. Also, $\langle\rangle$ is introduced to denote the densities obtained by conditioning on $\boldsymbol{\psi}$, $\langle\cdot|\cdot\rangle = [\cdot|\cdot, \boldsymbol{\psi}]$. All this simplified notation enables us to fit many equations in one line in the later chapters.

The Kalman filter and smoother can be used to calculate the likelihood $[\mathbf{y}_{1:T}|\boldsymbol{\psi}]$ and the posterior $[\boldsymbol{\theta}_{1:T}|\mathbf{y}_{1:T}, \boldsymbol{\psi}]$ ($\langle\boldsymbol{\theta}_{1:T}|\mathbf{y}_{1:T}\rangle$ in our simplified notation). The Kalman filter starts with the filtering density, where "filtering" means predicting the state $\boldsymbol{\theta}_t$ given the current observations $\mathbf{y}_{1:t}$. This is done recursively based on the previous

filtering density. We denote the previous filtering density at time $t-1$ as

$$\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1} \sim \mathrm{N}(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}), \tag{3.16}$$

for $t > 0$. At $t = 1$, it is equivalent to $\boldsymbol{\theta}_0 \sim \mathrm{N}(\mathbf{m}_0, \mathbf{C}_0)$, which initializes the Kalman filter. For $t > 1$, the filtering densities can be derived as follows.

With the state model (3.15), the one step forward filtering density is

$$\langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}\rangle = \int \langle \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\rangle \langle \boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}\rangle d\boldsymbol{\theta}_{t-1}. \tag{3.17}$$

It corresponds to the following Gaussian density

$$\begin{aligned}
&\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1} \sim \mathrm{N}(\mathbf{a}_t, \mathbf{R}_t) \\
&\mathbf{a}_t = \mathrm{E}(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) = \mathbf{G}_t\mathbf{m}_{t-1} &(3.18)\\
&\mathbf{R}_t = \mathrm{Var}(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T + \mathbf{U}_t. &(3.19)
\end{aligned}$$

The expression of the conditional mean (3.18) can be proved as

$$\mathbf{a}_t = \mathrm{E}\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}\right) = \mathrm{E}\left(\mathrm{E}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{y}_{1:t-1})\right) = \mathrm{E}\left(\mathbf{G}_t\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}\right) = \mathbf{G}_t\mathbf{m}_{t-1}.$$

Similarly, the conditional variance (3.19) can be proved as

$$\begin{aligned}
\mathbf{R}_t &= \mathrm{Var}(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) = \mathrm{Var}\left(\mathrm{E}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{y}_{1:t-1})\right) + \mathrm{E}\left(\mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{y}_{1:t-1})\right) \\
&= \mathrm{Var}(\mathbf{G}_t\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}) + \mathbf{U}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T + \mathbf{U}_t.
\end{aligned}$$

Considering the observation model (3.14), we can also easily derive the one step forward predictive density of $\mathbf{y}_t$

$$\langle \mathbf{y}_t|\mathbf{y}_{1:t-1}\rangle = \int \langle \mathbf{y}_t|\boldsymbol{\theta}_t\rangle \langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}\rangle d\boldsymbol{\theta}_t. \tag{3.20}$$

It corresponds to the following Gaussian density

$$\begin{aligned}
\mathbf{y}_t|\mathbf{y}_{1:t-1} &\sim \mathrm{N}(\mathbf{f}_t, \mathbf{Q}_t) \\
\mathbf{f}_t &= \mathrm{E}(\mathbf{y}_{t-1}|\mathbf{y}_{1:t-1}) = \mathbf{F}_t\mathbf{a}_t \\
\mathbf{Q}_t &= \mathrm{Var}(\mathbf{y}_{t-1}|\mathbf{y}_{1:t-1}) = \mathbf{F}_t\mathbf{R}_t\mathbf{F}_t^T + \mathbf{V}_t.
\end{aligned}$$

The above quantities can be derived similarly as (3.18) and (3.19). We will skip the details here and the reader can refer to Petris et al. (2009) for a complete treatment.

Notice with these predictive densities, we can calculate the likelihood of the SSM,

$$[\mathbf{y}_{1:T}|\boldsymbol{\theta}] = \langle \mathbf{y}_{1:T} \rangle = \prod_{t=1}^{T} \langle \mathbf{y}_t|\mathbf{y}_{1:t-1} \rangle.$$

Using Bayes rule and conditional independence property of the SSM, the current filtering density equals

$$\langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t} \rangle = \frac{\langle \mathbf{y}_t|\boldsymbol{\theta}_t \rangle \langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t-1} \rangle}{\langle \mathbf{y}_t|\mathbf{y}_{1:t-1} \rangle}, \tag{3.21}$$

which depends on the observation model (3.14), the one step forward filtering density (3.17), and the predictive density (3.20). As all three are Gaussian, the filtering density is also Gaussian,

$$\begin{aligned}
\boldsymbol{\theta}_t|\mathbf{y}_{1:t} &\sim \mathrm{N}(\mathbf{m}_t, \mathbf{C}_t) \\
\mathbf{m}_t &= \mathrm{E}(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) = \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t^T\mathbf{Q}_t^{-1}(\mathbf{y}_t - \mathbf{f}_t) \\
\mathbf{C}_t &= \mathrm{Var}(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) = \mathbf{R}_t - \mathbf{R}_t\mathbf{F}_t^T\mathbf{Q}_t^{-1}\mathbf{F}_t\mathbf{R}_t.
\end{aligned}$$

So far, we have shown how to derive the current filtering density $\langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t} \rangle$ based on the previous filtering density $\langle \boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1} \rangle$. An intermediate step (3.20) helps us calculate the likelihood of the DLM. This procedure is referred to as the Kalman filter algorithm (Kalman, 1960) and it utilizes the Gaussian Markovian property of the DLM to sequentialize the calculation. We are only moving forward in time with the Kalman filter. To calculate the posterior of $\boldsymbol{\theta}_{1:T}$ given all the observations, we also need to move backward in time. The corresponding procedure is usually

referred as the Kalman smoother.

Similarly as in the Kalman filter, we first set up a backward transition density

$$\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T} \sim \mathrm{N}(\mathbf{s}_{t+1}, \mathbf{S}_{t+1}).$$

At $t = T-1$, it is equivalent to the filtering density at the last time point $\boldsymbol{\theta}_T|\mathbf{y}_{1:T} \sim \mathrm{N}(\mathbf{m}_T, \mathbf{C}_T)$. With the Markovian property, we can show that

$$\langle \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:T}\rangle = \langle \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:t}\rangle = \frac{\langle \boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\rangle\langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t}\rangle}{\langle \boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t}\rangle}.$$

It helps us to further derive the following smoothing density

$$\langle \boldsymbol{\theta}_t|\mathbf{y}_{1:T}\rangle = \langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t}\rangle \int \frac{\langle \boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\rangle}{\langle \boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t}\rangle}\langle \boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T}\rangle d\boldsymbol{\theta}_{t+1}.$$

It depends on the filtering density $\langle \boldsymbol{\theta}_t|\mathbf{y}_{1:t}\rangle$ (3.21), the state model $\langle \boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\rangle$ (3.15), the one step forward predictive density $\langle \boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t}\rangle$ (3.17), and the previous smoothing density $\langle \boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T}\rangle$. All these densities are Gaussian and we can derive the smoothing density as

$$\begin{aligned}
&\boldsymbol{\theta}_t|\mathbf{y}_{1:T} \sim \mathrm{N}(\mathbf{s}_t, \mathbf{S}_t)\\
&\mathbf{s}_t = \mathbf{m}_t + \mathbf{C}_t\mathbf{G}_t^T\mathbf{R}_{t+1}^{-1}(\mathbf{s}_{t+1} - \mathbf{a}_{t+1})\\
&\mathbf{S}_t = \mathbf{C}_t - \mathbf{C}_t\mathbf{G}_{t+1}^T\mathbf{R}_{t+1}^{-1}(\mathbf{R}_{t+1} - \mathbf{S}_{t+1})\mathbf{R}_{t+1}^{-1}\mathbf{G}_{t+1}\mathbf{C}_t,
\end{aligned}$$

which is how the Kalman smoother calculates the current smoothing density $\langle \boldsymbol{\theta}_t|\mathbf{y}_{1:T}\rangle$ based on the previous smoothing density $\langle \boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T}\rangle$.

One key advantage of the Kalman filter and smoother is that they transfer the joint density of $\mathbf{y}_{1:T}$ and $\boldsymbol{\theta}_{1:T}$ into the sequential conditioning densities, such that the likelihood and posterior can be evaluated in linear time of T. It also pairs well with the conditional specification of our spline based SDE, as the SDE only specifies the one time step ahead conditional density instead of the joint density.

3.4 B-spline

The B-spline (or basis spline) is a popular class of basis functions used to approximate a non-periodic smooth function (see e.g., De Boor et al., 1978; Ramsay et al., 2009; Buderman et al., 2015). The B-spline basis functions are piecewise polynomials of order k that are continuous at the knots. Order k is one plus the degree of the polynomial, e.g., a constant function is of order one and a linear function is of order 2, etc. In addition, the B-spline has continuous derivatives up to the $(k-2)^{th}$ derivative. Given knots $t_0 < t_1 < \ldots < t_n$ and a desired order k, the B-spline construct $J = n+k-1$ basis functions recursively from $k = 1$, i.e., piecewise constant.

$$B_{j,1}(t) = \begin{cases} 1 & t_j \leq t < t_{j+1}, \\ 0 & \text{Otherwise} \end{cases}.$$

For order $k > 1$, the basis functions are constructed as follows:

$$B_{j,k}(t) = \frac{t - t_j}{t_{j+k-1} - t_j} B_{j,k-1}(t) + \frac{t_{j+k} - t}{t_{j+k} - t_{j+1}} B_{j+1,k-1}(t).$$

This recursive construction is implemented in the De Boor algorithm (see e.g., De Boor et al., 1978), which makes the B-spline very efficient to evaluate. Notice that we need to extend the knots beyond the t_0, t_n for the boundary cases and readers can refer to De Boor et al. (1978); Ramsay et al. (2009) for more detailed description. The knots $t_0, t_1, \ldots, t_n$ can have irregular spacing, but we only work with equally spaced knots in this thesis. We specify a set of B-spline basis functions via its number of basis functions J and order k, i.e., B-spline(J,k) hereafter.

The first derivative of the basis function $B_{j,k}$ with $k > 1$ is a linear combination of two basis functions from a set of B-splines with the same knots but order $k-1$.

$$B'_{j,k}(t) = \frac{k-1}{t_{j+k-1} - t_j} B_{j,k-1}(t) - \frac{k-1}{t_{j+k} - t_{j+1}} B_{j,k-1}(t).$$

Recall that $J = n+k-1$. This indicates that the first derivative of a B-spline with J basis functions and order k is formed by a B-spline with $J-1$ basis functions and order $k-1$. This property will be used later in Chapter 6.

Chapter 4

Bayesian Data Fusion Approaches

In this chapter, we propose two Bayesian data fusion approaches to combine tracking data from different sources.

4.1 Introduction

The problem of combining disparate data sets is not new to statisticians. In environmental statistics, for example, various approaches have been developed to combine measurements with numerical (computer) model outputs. Two of the most common approaches are Bayesian Melding (BM, Fuentes and Raftery, 2005) and downscaling (Berrocal et al., 2010; Zidek et al., 2012). Bayesian Melding was developed to combine direct observations of air–pollutant concentrations from a sparse network of monitoring stations with outputs by grid cell from a deterministic chemical transportation (computer) model in a geographical domain based on known pollutant source and geophysical information. In this approach, the direct observations and the computer model outputs are connected via a hidden process of the "true" air pollution level, that is, the monitoring stations observations $Z_0(s)$ at location s measure the true air pollutant level $Z(s)$ with some measurement error

$$Z_0(s) = Z(s) + e(s), \text{ for all } s,$$

where $e(s) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_e^2)$ is a white noise process. Notation $\cdot \overset{\text{iid}}{\sim} \cdot$ means the random variables on the left hand side are independently and identically distributed from a certain distribution on the right hand side. In addition, we assume the error process $\{\varepsilon(\cdot)\}$ is independent of $\{Z(\cdot)\}$. The computer model outputs $Z_1(s)$ are assumed to be

$$Z_1(s) = a(s) + b(s)Z(s) + \delta(s),$$

where $a(s)$ is the systematic additive error, $b(s)$ is the multiplicative error, and $\delta(s)$ is the additional noise. Usually, the multiplicative error $\{b(\cdot)\}$ is assumed to be a constant, and the additive error $\{a(\cdot)\}$ is modeled with polynomial or a Gaussian process with exponential covariance function (Foley and Fuentes, 2008; Fuentes and Raftery, 2005; Sahu et al., 2010). The BM approach has been adapted for other uses, such as to model hurricane surface winds (Foley and Fuentes, 2008), ozone levels (Liu et al., 2011), and wet deposition (Sahu et al., 2010), etc. These applications have demonstrated the remarkable flexibility and effectiveness of the BM approach.

Despite its flexibility, the BM approach is computationally cumbersome when dealing with "change-of-support" problems (Berrocal et al., 2010; Zidek et al., 2012)—namely where the observations are collected on different spatial scales, such as in-situ station measurements versus area averages. In response, Berrocal et al. (2010) developed a spatial or spatio–temporal downscaling approach to combine the air pollutant data, using a linear mixed effect model. This downscaling approach can be described, using the same notation as above, by the following regression model with spatially or temporally correlated coefficients,

$$Z_0(s) = \beta_0 + r_1(s) + (\beta_1 + r_2(s))Z_1(s) + \varepsilon(s),$$

where $\{r_1(\cdot)\}$ and $\{r_2(\cdot)\}$ are some spatio–temporal random processes, such as the Matérn process. Note that the index s can be time, location, or their combination. Recently, Rundel et al. (2015) further developed the downscaling approach to combine speciated $\mathrm{PM}_{2.5}$ (particulate matter with a diameter of 2.5 micro-meters or less) levels from multiple monitoring networks and computer model outputs.

In this paper, we adapt both the BM and downscaling approaches to combine multiple sources of observations for tracking objects. We use both approaches to combine the GPS locations and Dead-Reckoned paths of marine mammals and apply them to data from northern fur seals—a species that inhabits the North Pacific Ocean (Battaile et al., 2015). Unlike combining the observations from monitoring networks and computer model outputs, we are less concerned with the spatial "change-of-support" problem in tracking, as the observations and model outputs in our application lie on the same time scale.

In the BM framework, we first choose a random process that reflects the nature of the tracked object or the physics of its evolution, as the prior for $\{Z(\cdot)\}$. For example, we consider $\{Z(\cdot)\}$ as a Brownian Bridge process in our application to northern fur seal tracking, which corresponds to the fact that they return to their breeding beaches to feed their young after a foraging trip. To track an infectious disease, we could model $\{Z(\cdot)\}$ with the susceptible–exposed–infected–recovered (SEIR) compartmental equation (Dukic et al., 2012). All the systems of observations are linked to different transformations of $\{Z(\cdot)\}$: the direct observation $\{Z_0(\cdot)\}$ is $\{Z(\cdot)\}$ plus a white noise process while the indirect observation $\{Z_1(\cdot)\},\{Z_2(\cdot)\},\ldots$ are functions of $\{Z(\cdot)\}$ plus some other random processes as biases that reflect model error. Our BM framework can be summarized as follows: for all t,

$$\begin{aligned} Z(t) &\sim \text{A certain random process} \\ Z_0(t) &= Z(t) + \varepsilon_0(t) \\ Z_1(t) &= g_1(Z(t)) + \xi_1(t) \\ Z_2(t) &= g_2(Z(t)) + \xi_2(t) \\ &\cdots\cdots, \end{aligned} \tag{4.1}$$

where $\{\varepsilon_0(\cdot)\}$ is a white noise process. Functions $g_j(\cdot), j = 1,2,\ldots$ can be assumed to have certain parameterized form but with unknown parameters. The $\{\xi_j(\cdot)\}, j = 1,2,\ldots$ are random processes that representing the errors in the indirect observations. To make inference about $\{Z(\cdot)\}$, we need to calculate the posterior distribution of $\{Z(\cdot)\}|\{Z_0(\cdot)\},\{Z_1(\cdot)\},\{Z_2(\cdot)\},\ldots$, whose posterior mean can be

the smoothed/predicted "track" of the tracking object and the posterior credible intervals (CI) reflect the uncertainty in the track. Note that the random process in (4.1), such as the Brownian Bridge or the SEIR process, only reflects our prior knowledge of the track. Its posterior is updated via the observations.

The downscaling approach for tracking bypasses the modeling of $\{Z(\cdot)\}$ and builds the mixed effects model between the direct observations and the other systems of observation as

$$Z_0(t) = \beta_0 + r_0(t) + (\beta_1 + r_1(t))\, Z_1(t) + (\beta_2 + r_2(t))\, Z_2(t) + \cdots + \varepsilon(t),$$

for all t. $\{r_j(\cdot)\}, j = 0, 1, 2, \ldots$, can be Gaussian processes as in Berrocal et al. (2010) and Zidek et al. (2012). As with the BM approach, the posterior mean and CI of the linear predictor can be the predicted "track" and its uncertainty.

For the tracking application, we need to first choose appropriate processes for the random components, such as $\{Z(\cdot)\}, \{\xi_j(\cdot)\}$, etc. Besides matching the physics of the tracked objects, we also need to take account of the computational burden. For example, devices attached to a northern fur seal, which samples at 16Hz, and the video tracking of NBA players (Liu et al., 2016), which samples at 25Hz, both yield incredibly big data sets. As a result, we avoid using MCMC techniques that have been used in the past for both the BM and downscaling approaches. Instead, we fit the downscaling model with the integrated nested Laplace approximation (INLA) method developed in Rue et al. (2009) and Lindgren et al. (2011). Inspired by the sparse matrix techniques, likelihood approximations, and gradient based numeric integrations in the INLA approach, we exploit the properties of the processes and designed approximations to the likelihood for the BM approach.

The following describes the BM and downscaling approaches for tracking, and applies them to a case study of northern fur seals as reviewed in Chapter 2. Section 4.2 describes our Bayesian Melding approach while Section 4.3 describes the downscaling approach. We perform several simulation studies to evaluate our BM and downscaling approaches, which are reported in Section 4.4. Section 4.5 contains the case study results together with cross–validation comparisons. Conclusion and discussion are contained in Section 4.6.

4.2 Bayesian melding

In this section, we introduce the BM approach to combine the information from the accurate but sparse GPS observations with the biased but dense DR path. For simplicity, the two dimensions of the path (latitude and longitude) are dealt with separately. Abstractly, our theory is about a one dimensional path over time, which we denote by $\eta(t)$ at discrete time points $t = 0, 1, 2, \ldots, T$. The time unit plays no essential role in our theory. The approach works just as well with unequally spaced time points, i.e., for arbitrary $t_0, t_1, \ldots, t_T$. But for expository simplicity we work with $0:T$ because the DR path is equally spaced. As in the previous BM literature, we put a Gaussian process prior on $\eta(t)$,

$$\boldsymbol{\eta}(0:T) \sim N\left(\mathbf{f}(0:T), \mathbf{R}(0:T, 0:T)\right), \tag{4.2}$$

where $\mathbf{f}(\cdot)$ denotes the process mean function and $\mathbf{R}$, its covariance matrix. $0:T$ denotes $0, 1, 2, \ldots, T$. Similarly, $\mathbf{f}(0:T)$ stands for the vector $(f(0), f(1), \ldots, f(T))^T$ while $\mathbf{R}(0:T, 0:T)$ is a $(T+1) \times (T+1)$ covariance matrix whose (t, t') entry $R(t, t')$ equals $\mathrm{Cov}\left(\eta(t), \eta(t')\right)$. Throughout this thesis, bold faced characters are used exclusively to represent vectors or matrices.

Various options are available for this Gaussian process. A common one (Fuentes and Raftery, 2005; Sacks et al., 1989) assumes that $\mathbf{f}$ is a simple parametric model, e.g., a constant or a linear function of the index (time or location) as well as additional covariates that may affect $\boldsymbol{\eta}$. The covariance function is usually assumed as $\sigma^2 \rho(|t - t'|)$, where $\rho(\cdot)$ is an isotropic correlation function such as the Matérn or power exponential. However, this popular stationary Gaussian process is not suitable for our application. As noted above, the tracked animal must return to the starting location of its foraging trip (to reunite with her pup), which means that the start and end points of the track are fixed, as shown in Figure 4.1. Apart from the start and end points, the animal's path is unknown, and hence random in our Bayesian framework. Its variation is relatively large in the middle and small when close to the known start and end points. These features of the path led us to model

it with a Brownian Bridge process, whose mean and covariance functions are:

$$\begin{aligned} f(t) =& A + (B-A)\frac{t}{T} \\ R(s,t) =& \sigma_H^2 \frac{(\min(s,t))(T-\max(s,t))}{T} \end{aligned}$$

where $\eta(0) = A$ and $\eta(T) = B$ are the known start and end points of the path, while σ_H^2 is the variance parameter. Notice that $R(0,\cdot) = R(\cdot,T) = 0$, in accordance with the known start and end points $\eta(0)$ and $\eta(T)$. Also, $R(t,t)$ increases with t when $t < T/2$ and decreases with t for $t > T/2$, reflecting the fact that the variation of the path is largest in the middle. Another noteworthy property of our covariance matrix $\mathbf{R}$ is its form as the product of a scalar σ^2 and a matrix, the latter depending only on the time points. To clearly represent the parameters of the Brownian Bridge process, we introduce the notation

$$\mathrm{BB}(A, B, T_S, T_E, \sigma^2) \tag{4.3}$$

for a Brownian Bridge process, which starts from A at time T_S and ends in B at time T_E with a variance parameter σ^2, namely having mean function $f(t) = A + (B-A)(t-T_S)/(T_E-T_S)$ and covariance function $\sigma^2(\min(s,t)-T_S)(T_E - \max(s,t))/((T_E-T_S))$.

Our choice of the Brownian Bridge prior is popular in the literature of biology and ecology. According to Humphries et al. (2010), marine mammals tend to exhibit Brownian–like movements in environments with abundant food resources, such as the resource filled ocean around Bogoslof island where our case study was centered (Benoit-Bird et al., 2013b). Also, a Brownian Bridge model was proposed by Horne et al. (2007) to model the habitat usage of a wide range of animal species. This model was further improved by Kranstauber et al. (2012, 2014), and Sawyer et al. (2009). Pozdnyakov et al. (2014) studied different ways of estimating the parameters of the Brownian Bridge model. Many other examples where animal paths have been modeled with Brownian Bridge processes can be found in the references of the above papers.

It should be recognized that the Brownian Bridge prior does not mean that the animal's path after being updated with the GPS observations and the DR path is still

a Brownian Bridge. We use the Brownian Bridge prior to motivate a proper covariance structure, whereby the beginning and end of the animal's path are known but there is more uncertainty about the middle part of this path. The Markovian structure of the Brownian Bridge also helps to simplify the Bayesian computation of combining this process with the DR path, as discussed later.

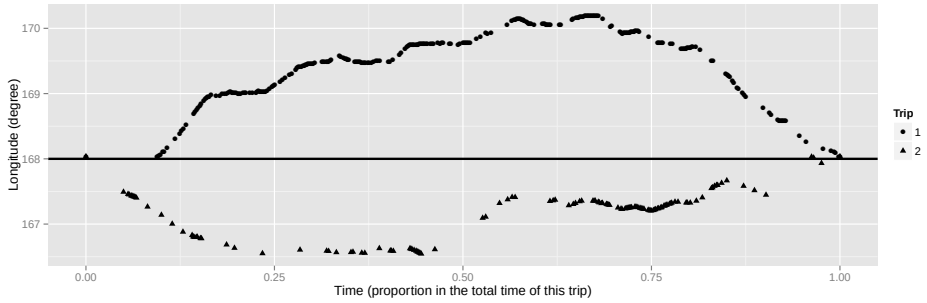


Figure 4.1: The (negative) longitude of the GPS observations of the foraging trips made by two northern fur seals in our case study. Both trips started and ended at Bogoslof Island (Alaska) where the females gave birth and nursed their pups, and the horizontal line indicates the longitude -168.035E of the island. The time unit is the proportion of the total time of this foraging trip.

The GPS observations are available at time points $t_0 = 0 < t_1 < t_2 < \ldots < t_k = T$, where the t_k's form a subset of $\{0,1,\ldots,T\}$. The GPS observations are further denoted by $Y(t_k), k = 0,1,\ldots,K$. They are unbiased observations of the true location:

$$Y(t_k)|\eta(t_k) \overset{\text{iid}}{\sim} \mathrm{N}(\eta(t_k), \sigma_G^2), \tag{4.4}$$

for $k = 1,\ldots,K-1$. The known start and end points assumption implies that $Y(t_0) = \eta(t_0) = A$, and $Y(t_K) = \eta(t_K) = B$ are known.

Next, $X(t), t = 0,1,2,\ldots,T$ is used to denote the observed DR path. To incorporate the bias of the DR path, we assume:

$$X(t) = \eta(t) + h(t) + \xi(t), \tag{4.5}$$

where $h(\cdot;\boldsymbol{\beta})$ is a parametric function designed to capture the systematic bias trend in the DR path. $\{\xi(\cdot)\}$ denotes another Gaussian process independent of $\{\eta(\cdot)\}$ that captures any irregular components in the deviation of the DR path from the truth:

$$\boldsymbol{\xi}(1:T) \sim \mathrm{N}(\mathbf{0}, \mathbf{C}(1:T, 1:T)).$$

For the parametric bias component, we have considered various models, e.g., $h(t) = \sum_{i=1}^{Q} \beta_i t^{i-1}$. The bias process $\{\xi(\cdot)\}$ is assumed to be a Brownian motion process (random walk of order 1) whose covariance function is therefore

$$C(s,t) = \sigma_D^2 \min(s,t).$$

We believe the Brownian motion process to be a reasonable approximation to the gradually accumulating error in the DRA. If we assume the errors in the velocity estimates from the DRA, after removing the systematic bias $h(t)$, at each time point are an $i.i.d.$ normal sequence, the error in the integrated path is then a Brownian motion.

The final ingredients in our BM model are the prior distributions of the parameters. For notational simplicity, all densities are denoted by square brackets $[\ldots]$ throughout this thesis. For σ_G^2, we assume a known constant based on the previous extensive tests of the Fastloc® GPS device (Bryant, 2007). The priors of the other two variance parameters are chosen to be $[\sigma_H^2] \propto 1/\sigma_H^2$ and $[\sigma_D^2] \propto 1/\sigma_D^2$, which are uniform priors on the log scale ($[\log(\sigma_H^2)] \propto 1$). For $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_Q)^T$, a flat prior $[\boldsymbol{\beta}] \propto 1$ is used. All these parameters are assumed to be independent of each other in their priors.

For expository simplicity in describing the joint distribution of all the data and parameters, the following notation is introduced:

- The unknown part of the true path is denoted by $\boldsymbol{\eta} = \boldsymbol{\eta}(1:(T-1))$, a $T-1$ dimensional vector.

- The GPS observations of the unknown part of the path are denoted by $\mathbf{Y} = (Y(t_1), Y(t_2), \ldots, Y(t_{K-1}))^T$, a $K-1$ dimensional vector.

- The DR path is $\mathbf{X} = \left(\mathbf{X}(1:(T-1))^T, X(T) - Y(T)\right)^T$, a vector of dimension T.
- For the two unknown variance parameters, let $\boldsymbol{\phi} = (\sigma_H^2, \sigma_D^2)^T$.

The joint likelihood of our model is

$$[\mathbf{X}, \mathbf{Y}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\phi}] = [\boldsymbol{\phi}]\,[\boldsymbol{\beta}]\,[\boldsymbol{\eta}|\boldsymbol{\phi}]\,[\mathbf{Y}|\boldsymbol{\eta}]\,[\mathbf{X}|\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\eta}]. \tag{4.6}$$

To obtain an estimate of the animal's true path and its uncertainty, we need the posterior distribution

$$[\boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}] = \int \underbrace{[\boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\phi}]}_{(1)} \times \underbrace{[\boldsymbol{\phi}|\mathbf{X}, \mathbf{Y}]}_{(2)} \mathbf{d}\boldsymbol{\phi}. \tag{4.7}$$

We also include the $\boldsymbol{\beta}$ term, which can be used to assess the bias of the DRA. The posterior mean, denoted by $\tilde{\eta}(t)$, can be an estimate of the animal's path and the posterior standard error, denoted by $\tilde{\sigma}(t)$ provides an uncertainty statement about the estimated path. The point-wise 95% credible interval for $\eta(t)$ can be constructed via a normal approximation

$$[\tilde{\eta}(t) - 1.96\tilde{\sigma}(t), \tilde{\eta}(t) + 1.96\tilde{\sigma}(t)]. \tag{4.8}$$

In principle, the "exact" credible intervals can be found via the normal mixtures in our numerical integration scheme. Yet, we found empirically in our case study that the exact credible intervals were almost indistinguishable from the normal approximated intervals in (4.8). This is discussed in detail in Section 4.5.1.

4.2.1 Model inference

To calculate the posterior (4.7), we first fix the variance parameters $\boldsymbol{\phi}$ and calculate part (1) in Equation (4.7) and then integrate over the posterior of $\boldsymbol{\phi}$. The first part of this section shows how the components of (4.7) can be efficiently evaluated. We then use numerical integration on an adaptive grid, as in INLA (Rue et al., 2009), to marginalize over the randomness in $\boldsymbol{\phi}$. The numerical integration part is described in Appendix A.6.

For notational simplicity, $\langle\cdot|\cdot\rangle$ denotes $[\cdot|\cdot,\boldsymbol{\phi}]$, that is $\langle\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}\rangle = [\boldsymbol{\eta}|\mathbf{X},\mathbf{Y},\boldsymbol{\phi}]$. As we specify our model in a Gaussian and linear fashion, it is straightforward to show that $\langle\boldsymbol{\beta},\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}\rangle$ is a multivariate Gaussian density,

$$\langle\boldsymbol{\beta},\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}\rangle \propto \exp\left\{-\frac{1}{2}\left((\boldsymbol{\zeta}-\mathbf{M}_1^{-1}\mathbf{M}_2)^T\mathbf{M}_1(\boldsymbol{\zeta}-\mathbf{M}_1^{-1}\mathbf{M}_2)\right)\right\} \tag{4.9}$$

where $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T,\boldsymbol{\eta}^T)^T$ and $\mathbf{M}_1,\mathbf{M}_2$ are derived in Appendix A.1.

Although the multivariate Gaussian posterior makes inference conceptually easy in implementation, calculating its posterior mean $\mathbf{M}_1^{-1}\mathbf{M}_2$ and covariance matrix $\mathbf{M}_1^{-1}$ actually involves a matrix decomposition with computational complexity of order $O(T^3)$, which is a tremendous computational burden when T is large. It is possible to avoid the $O(T^3)$ matrix decomposition with certain sparse matrix techniques together with the Sherman–Morrison–Woodbury formula (Henderson and Searle, 1981), but those techniques still require the storage of huge matrices and complicated matrix calculations. This pushes us to further reduce the complexity of (4.9).

It is easily seen that we have more information (data) about $\boldsymbol{\eta}_G \triangleq \boldsymbol{\eta}(t_{0:K})$ where the GPS observations are available than where they are not. For $\boldsymbol{\eta}(0:T\setminus t_{0:K})$, we only have the DR path. So our first step breaks $\boldsymbol{\eta}$ into two sets, that is

$$\langle\boldsymbol{\beta},\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}\rangle = \langle\boldsymbol{\eta}(0:T\setminus t_{0:K})|\boldsymbol{\beta},\boldsymbol{\eta}_G,\mathbf{X},\mathbf{Y}\rangle\,\langle\boldsymbol{\beta},\boldsymbol{\eta}_G|\mathbf{X},\mathbf{Y}\rangle. \tag{4.10}$$

We can then use the Markovian property of the Brownian Bridge process (see e.g., Stirzaker and Grimmett, 2001) to simplify (4.10) as:

$$\begin{aligned}\langle\boldsymbol{\beta},\boldsymbol{\eta}|\mathbf{X},\mathbf{Y},\boldsymbol{\phi}\rangle = &\left\{\prod_{k=0}^{K-1}\langle\boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},\mathbf{X},\mathbf{Y}\rangle\right\}\\ &\times\langle\boldsymbol{\beta},\boldsymbol{\eta}_G|\mathbf{X},\mathbf{Y}\rangle.\end{aligned} \tag{4.11}$$

A hidden assumption in (4.11) is that $t_{k+1} > t_k+1$ for all k, such that each period has at least one non-GPS observations. When $t_{k+1} = t_k+1$, our implementation skips this period automatically.

In this way, we partition the long $\boldsymbol{\eta}$ series into small pieces separated by the

GPS observations. The next step exploits the Markovian property of the Brownian Motion and enables us to simplify the k^{th} term in the first part of (4.11) as

$$\begin{aligned}&\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},\mathbf{X},\mathbf{Y}\rangle = \\ &\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},\mathbf{X}(t_k:t_{k+1})\rangle. \end{aligned} \tag{4.12}$$

All the derivations for (4.11) and (4.12) are provided in Appendix A.2. In (4.12), the posterior of $\eta(t)$ between two GPS points can be evaluated using the corresponding DR path together with the posterior distribution of the two GPS points and $\boldsymbol{\beta}$. This remarkably reduces the memory cost when computing the posterior of the long sequence and enables us to easily parallelize the whole calculation. Moreover, both the Brownian Bridge and Brownian Motion processes conditioned on two end points are Brownian Bridge processes, such that,

$$\begin{aligned}\boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}) &\sim \text{BB}(\eta(t_k),\eta(t_{k+1}),t_k,t_{k+1},\sigma_H^2)\\ \boldsymbol{\xi}(t_k+1:t_{k+1}-1)|\xi(t_k),\xi(t_{k+1}) &\sim \text{BB}(\xi(t_k),\xi(t_{k+1}),t_k,t_{k+1},\sigma_D^2).\end{aligned}$$

This fact is exploited to completely avoid the matrix inverse calculation when evaluating (4.12), which further reduces the computational burden. The derivations are included in Appendix A.3. Also, we found that the bias correction in the most simplified BM approach (empirical Bayesian, $\boldsymbol{\beta}=\mathbf{0}$) is a shrinkage of the conventional bias correction, which will account for the signal-noise ratio in the DR path. This is discussed in Section 4.2.2.

However, the evaluation of $\langle \boldsymbol{\beta},\boldsymbol{\eta}_G|\mathbf{X},\mathbf{Y}\rangle$ in (4.11) still requires us to deal with the long sequence $\mathbf{X}$. But $\mathbf{Y}$ is an unbiased observation of $\boldsymbol{\eta}_G$ and therefore $\langle \boldsymbol{\eta}_G|\mathbf{X},\mathbf{Y}\rangle$ can be well approximated by $\langle \boldsymbol{\eta}_G|\mathbf{Y}\rangle$. This approximation is exceptionally good when $\sigma_D^2 > \sigma_G^2$. The coefficients $\boldsymbol{\beta}$ can be well inferred from the difference between $\mathbf{X}_G \triangleq \mathbf{X}(t_{1:K})$ and $\mathbf{Y}$. Thus, we introduce the following approximation:

$$\langle \boldsymbol{\beta},\boldsymbol{\eta}_G|\mathbf{X},\mathbf{Y}\rangle \approx \langle \boldsymbol{\beta},\boldsymbol{\eta}_G|\mathbf{X}_G,\mathbf{Y}\rangle. \tag{4.13}$$

With similar arguments, we can also approximate the posterior of $\boldsymbol{\phi}$ by,

$$[\boldsymbol{\phi}|\mathbf{X},\mathbf{Y}] \approx [\boldsymbol{\phi}|\mathbf{X}_G,\mathbf{Y}] \tag{4.14}$$

The explicit expressions for (4.13) and (4.14) are included in Appendix A.4. Our simulations that mimic the real data sets have shown that the impact of the two approximation errors in (4.13) and (4.14) is negligible. We also verified through a thinned version of the real data set that the difference between the posterior obtained from (4.13) and (4.14) is not significant.

In summary, the posterior of $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ is approximated as follows:

$$\begin{aligned}
[\boldsymbol{\eta},\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}] &= \int [\boldsymbol{\eta},\boldsymbol{\beta}|\mathbf{X},\mathbf{Y},\boldsymbol{\phi}][\boldsymbol{\phi}|\mathbf{X},\mathbf{Y}]d\boldsymbol{\phi} \\
&= \int \langle \boldsymbol{\eta}(0:T\setminus t_{0:K})|\boldsymbol{\beta},\boldsymbol{\eta}_G,\mathbf{X},\mathbf{Y}\rangle \langle \boldsymbol{\eta}_G,\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}\rangle [\boldsymbol{\phi}|\mathbf{X},\mathbf{Y}]d\boldsymbol{\phi} \\
&= \int \left\{ \prod_{k=0}^{K-1} \langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\boldsymbol{\beta},\eta(t_k),\eta(t_{k+1}),\mathbf{X}\rangle \right\} \langle \boldsymbol{\eta}_G,\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}\rangle [\boldsymbol{\phi}|\mathbf{X},\mathbf{Y}]d\boldsymbol{\phi} \\
&= \int \left\{ \prod_{k=0}^{K-1} \langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\boldsymbol{\beta},\eta(t_k),\eta(t_{k+1}),\mathbf{X}(t_k:t_{k+1})\rangle \right\} \\
&\qquad \times \langle \boldsymbol{\eta}_G,\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}\rangle [\boldsymbol{\phi}|\mathbf{X},\mathbf{Y}]d\boldsymbol{\phi} \\
&\approx \int \left\{ \prod_{k=0}^{K-1} \langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\boldsymbol{\beta},\eta(t_k),\eta(t_{k+1}),\mathbf{X}(t_k:t_{k+1})\rangle \right\} \\
&\qquad \times \langle \boldsymbol{\eta}_G,\boldsymbol{\beta}|\mathbf{X}_G,\mathbf{Y}\rangle [\boldsymbol{\phi}|\mathbf{X}_G,\mathbf{Y}]d\boldsymbol{\phi}.
\end{aligned} \tag{4.15}$$

The integration in expression (4.15) is calculated on an adaptive grid based on $[\boldsymbol{\phi}|\mathbf{X}_G,\mathbf{Y}]$, which is discussed in Appendix A.6. Combining all these techniques to simplify computation reduces the computational time of our BM approach to a level similar to that of the DRA. For the two data sets we worked with, the DRA and BM of the DR path and GPS took less than five minutes in total on a regular laptop. We implemented the BM approach in an R package "BayesianAnimalTracker" (Liu, 2014). The speed with which calculations can be done using this package, i.e., Table 3 of Liu et al. (2015), is a huge benefit for marine biologists who want to follow the animal while aboard a ship or on a remote island using a

portable laptop without access to the Internet or any high performance computers.

4.2.2 Connection with the conventional correction

When compared to the conventional bias correction method in (2.1), our BM approach can account for both data and model uncertainty and provide a CI for the estimates of the animal's path. Moreover, there is an interesting connection between the BM posterior mean $\tilde{\eta}$ and the conventional corrected path $\hat{\eta}$ in Equation (2.1). Without loss of generality, let $K = 1$ (no GPS observations except the known start and end points) and $Y(0) = X(0) = 0$. So (2.1) can be written as

$$\hat{\eta}(t) = Y(T)\frac{t}{T} + \left(X(t) - X(T)\frac{t}{T}\right).$$

For BM, if $h(t) = 0$ for all t and $\boldsymbol{\phi}$ is known, the posterior mean $\tilde{\boldsymbol{\eta}}$ under the above assumptions as calculated by (A.15) in Appendix A.5 simplifies to

$$\tilde{\eta}(t) = Y(T)\frac{t}{T} + \frac{\sigma_H^2}{\sigma_H^2 + \sigma_D^2}\left(X(t) - X(T)\frac{t}{T}\right).$$

The first parts of $\hat{\eta}(t)$ and $\tilde{\eta}(t)$ represent linear interpolations between $Y(0) = 0$ and $Y(T)$, which determines the basic trend of the animal's path between two known points. The second part is a "bridge" constructed by the $X(t)$, which starts at $X(0) = 0$ and ends at $0 = X(T) - X(T)T/T$. This bridge can be treated as the "detail" for the animal's path, which is then added to the basic trend of the first part.

The difference between the traditional conventional method and the simplified BM approach is the weight on the "detail". In the conventional approach, the "detail" is directly added to the basic trend while BM shrinks the detail by a factor of $\rho = \sigma_H^2/(\sigma_H^2 + \sigma_D^2)$. According to our model, we cannot distinguish between $\eta(t)$ and $\xi(t)$ in $X(t)$ at those non-GPS points, as we only observe the sum of them, but we know that $\eta(t)$ accounts for the σ_H^2 part of the total $\sigma_H^2 + \sigma_D^2$ variance (they are both of mean zero after $Y(T)t/T$ is removed). In this way, a fraction $\rho = \sigma_H^2/(\sigma_H^2 + \sigma_D^2)$ of the detail is treated as signal in $\eta(t)$ and added to the basic linear trend.

Notice that we only compare the most simplified BM approach to the con-

ventional approach. In practice, the BM $\tilde{\eta}$ is far more complicated than the form shown above with the parametric part from $\boldsymbol{\beta}$ and the integration over $\boldsymbol{\phi}$. According to our simulations and cross-validation of real data later in this chapter, our $\tilde{\eta}$ is remarkably better than the conventional $\hat{\eta}$.

4.3 Downscaling

Using the same notation as in the previous section, we propose the following downscaling model for the GPS observations and DR path:

$$Y(t_k) = \beta_0 + \gamma_1(t_k) + (\beta_1 + \gamma_2(t_k))X(t_k) + \varepsilon(t_k), \tag{4.16}$$

where $\{\gamma_1(\cdot)\}$ and $\{\gamma_2(\cdot)\}$ are zero–mean Gaussian processes, such as random walks of order 1 and 2 (RW1, RW2), and autoregressive processes of order 1, 2, and 3 (AR1, AR2, AR3). $\{\varepsilon(\cdot)\}$ is a white–noise process. For expository simplicity, let $\boldsymbol{\gamma}$ denote all the unknown and random components in (4.16), including $\beta_0, \beta_1, \boldsymbol{\gamma}_1(0:T), \boldsymbol{\gamma}_2(0:T)$ and the hyper parameters governing them, e.g., the variance/correlation parameters of $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \varepsilon$. The combined path of the tracked animal can be learned from the posterior

$$\begin{aligned}
&[Y(t)|X(t), \mathbf{X}(t_{0:K}), \mathbf{Y}(t_{0:K})] = \int_{\boldsymbol{\gamma}} [Y(t), \boldsymbol{\gamma}|X(t), \mathbf{X}(t_{0:K}), \mathbf{Y}(t_{0:K})] d\boldsymbol{\gamma} \\
&= \int_{\boldsymbol{\gamma}} [Y(t)|\boldsymbol{\gamma}, X(t), \mathbf{X}(t_{0:K}), \mathbf{Y}(t_{0:K})] \, [\boldsymbol{\gamma}|\mathbf{X}(t_{0:K}), \mathbf{Y}(t_{0:K})] d\boldsymbol{\gamma},
\end{aligned}$$

for $t \in (0:T) \backslash t_{0:K}$. Notice that our formulation of the downscaling approach does not include a "true" process $\boldsymbol{\eta}$ explicitly as in the BM approach. Such an $\boldsymbol{\eta}$ can be calculated as $\eta(t) = \beta_0 + \gamma_1(t) + (\beta_1 + \gamma_2(t))X(t)$ for all t, which has the same posterior expectation as $Y(t)$.

Traditionally, the above integral has been calculated using an MCMC approach such as the Gibbs sampler (Berrocal et al., 2010; Zidek et al., 2012). However, the MCMC approach is computationally expensive as well as technically challenging and it sometimes encounters mixing or convergence problems. We therefore calculate the above posterior by the integrated-nested Laplace approximation (INLA) via the R-INLA package (Martins et al., 2013), as reviewed in Section 3.1.3.

For the downscaling approach, we use the default priors in R-INLA, which are all weakly informative priors. It is noteworthy that we do not put an informative prior on the variance parameter of $\varepsilon(t)$ as we do for the BM approach in Equation (4.4) because $\varepsilon(t)$ represents not only the measurement error in $Y(t)$, but also the lack-of-fit errors of the downscaling model.

So far, we have not found any direct equivalence between the BM and downscaling models, but notice that the posterior in both cases is essentially only calculated based on the DR path at the GPS observations $\mathbf{X}(t_{0:K})$, not on the full set $\mathbf{X}(0:T)$. This supports the use of our approximation to the likelihood in (4.13) and (4.14) from another perspective.

4.4 Simulation study

We performed several simulation studies to evaluate the performance of our approaches. These included a simulation to study the approximation in our BM approach (Section 4.4.1) and a comparison of all five approaches under three different data generating models (Section 4.4.2).

4.4.1 Simulation evaluation of the approximations in the BM approach

We used a simulation to evaluate the impact of our likelihood approximation in (4.13) and (4.14). For expository reasons, let "full BM" denote the Bayesian Melding approach based on the full likelihood (left hand side of (4.14)), and "approximate BM" denote Bayesian Melding approach based on the approximate likelihood (right hand side of (4.14)).

In this simulation, the data were generated according to our BM model: The true curve was simulated as a Brownian Bridge with $T = 2000$; the $K = 125$ GPS time points were randomly chosen from the $T = 2000$ time points; the GPS observations were $i.i.d.$ normal observations of the true curve at these time points; the DR path was the true curve plus the bias function $h(t)$ and another Brownian Motion process. The parameters used in the simulation were the maximum likelihood estimates from our Trip 2 Northing data set. The results shown below are based on 1000 replicates. Similar findings were found from other settings and thus omitted.

The first two panels in Figure 4.2, include the box–whisker plots of the posterior mode of $\log(\sigma_H^2)$ and $\log(\sigma_D^2)$ from the full and approximate BM. As we used a uniform prior on the log scale, the posterior modes are equivalent to the maximum likelihood estimates based on either the full or the approximate likelihood. Both estimates were nearly unbiased in both parameters, but the approximate likelihood estimates were obviously less efficient than the full likelihood estimates. The ratios between the standard errors of the approximate likelihood estimates and full likelihood estimates were 1.19 and 1.38 for $\log(\sigma_H^2)$ and $\log(\sigma_D^2)$ respectively. The increase in standard error, or equivalently the efficiency loss, was relatively small as the approximate BM only uses a $(125+125)/(2000+125) = 2/17$ fraction of the observations used in the full BM for parameter inference. For example, if we assume the standard error of the estimates was proportional to $1/\sqrt{n}$ as in the $i.i.d.$ case, the ratio between the standard errors would be predicted to be about $\sqrt{17/2} \approx 2.915$, far more than the ratio observed in the simulation study.

Moreover, the estimates of $\log(\sigma_H^2)$ and $\log(\sigma_D^2)$ are relatively unimportant, given that these are nuisance parameters in our Bayesian inference. What matters in the real application is the quality of the reconstructed path. Therefore, we calculated the root mean integrated squared error (RMISE, $\sqrt{\sum_{t=0}^{T}(\eta(t)-\bar{\eta}(t))^2/(T+1)}$) between the true curve $\eta(t)$ and different estimates $\bar{\eta}(t)$. $\bar{\eta}(t)$ can generically represent the posterior mean of $\eta(t)$ from either the full or approximate posterior, or the posterior mean of $Y(t)$ in the downscaling approach.

The third panel of Figure 4.2 is the scatter plot of the RMISE among all the replicates in our simulation. Clearly, most of the points lie on the diagonal line, indicating little difference between the reconstructed $\eta(t)$'s from approximate BM and full BM. This is because of the small efficiency loss in our likelihood approximation and the fact that we marginalize over the variance parameters in the posterior.

4.4.2 Comparison of all five approaches

The above simulation study suggests that our approximate BM inference procedure is quite accurate. In the following simulation, we compared all the five candidate methods of estimating the animal's path: the two methods working with the GPS

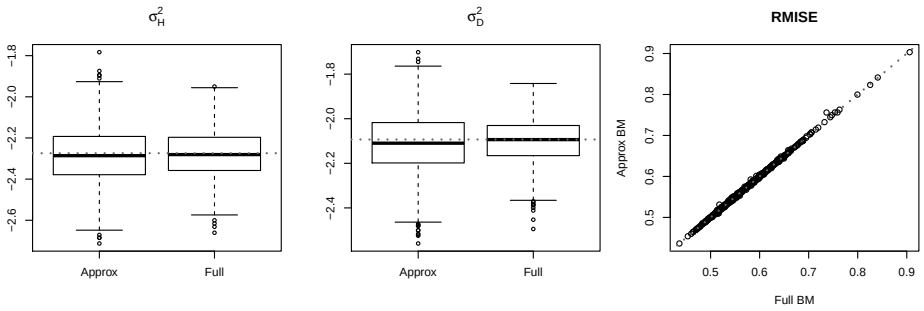


Figure 4.2: Plots from our simulation, from left to right: the box plot of the posterior mode of $\log(\sigma_H^2)$ obtained from the approximate and full likelihoods; the box plot of the posterior modes of $\log(\sigma_D^2)$; scatter plot of the RMISE of the posterior mean of η from the approximate Bayesian Melding approach versus the RMISE of the posterior mean from the full Bayesian Melding approach.

observations only (1. Linear interpolation and 2. CRAWL with drift term); and the three methods working with both the GPS and DR path (3. Conventional bias correction of the DR path, 4. Our Bayesian Melding method, and 5. Our downscaling method). We only work with the approximate BM because the full BM was very time consuming to run. The detailed model choices for the BM and downscaling were chosen to be the same as those selected for our real data application in the next section. That is, the BM model had $h(t)$ as a constant and the downscaling model had both γ_1 and γ_2 as RW1.

The data were generated from three models: the Bayesian Melding model as mentioned above, the CRAWL model, and the downscaling model. When generating data from the CRAWL model, a path was first simulated from a CRAWL model fitted to the real data sets and the DR path was generated by adding a Brownian Motion to this path. In the downscaling model, we fixed the observed DR path (from the real data), and generated the GPS observations by our downscaling model. The parameter settings of all the data generation models were the estimates from our real data from Trip 2. One thousand replicates were generated from each model.

To summarize the performance in both dimensions, we report the pooled RMISE

from both the northing and easting dimensions, namely,

$$\sqrt{\sum_{t=0}^{T}[(\eta_N(t)-\bar{\eta}_N(t))^2+(\eta_E(t)-\bar{\eta}_E(t))^2]/(2*(T+1))},$$

where N stands for Northing and E stands for Easting dimensions. The pooled RMISE serves as a general measure of the goodness of approximation in both Northing and Easting dimensions. The box-plots of the RMISE in the two dimensions separately were similar and thus omitted. The box–plots of the pooled RMISE are in Figure 4.3. Our BM and downscaling methods have smaller RMISEs than all the other competitors regardless of which model generated the real path. That is, even if the animal's path were generated from a CRAWL model (with an integrated OU process) instead of a Brownian Bridge, the BM approach with a mis–specified prior still manages to outperform CRAWL interpolation with the help from the DR path. In the comparison of BM and downscaling, their differences were very small. These findings were consistent with our findings in the cross–validation studies of the real data.

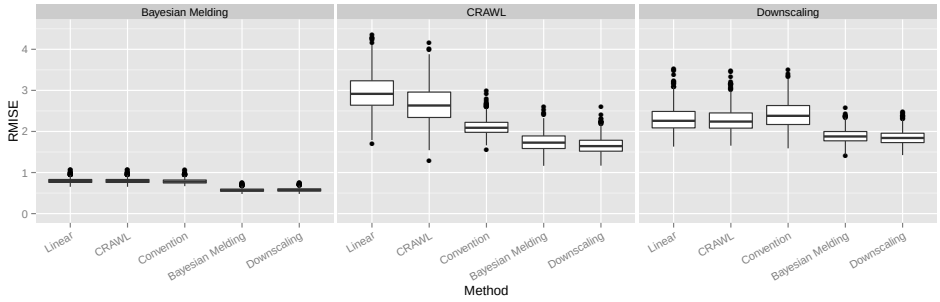


Figure 4.3: Pooled RMISE for all five methods of estimating animal's path, stratified by the data generating model.

Another interesting finding is that the conventional bias correction of the DR path outperformed CRAWL when the data were generated from the CRAWL model but not when the data were generated by the the downscaling model. This is because in the CRAWL data generation, the error in the DR path is purely additive while the error is both additive and multiplicative in the downscaling data generation. The conventional bias correction is only designed to deal with the additive

error but not the multiplicative one. On the other hand, as the model parameters in the BM approach are adaptive in accordance with the observations, it still manages to predict the animal's path as well as the correctly specified downscaling model.

4.5 Case study results

We used our proposed BM and downscaling approaches to combine the DR path and GPS observations for high resolution paths of northern fur seals. To simplify computation and comparing different models, we thinned the original 16Hz DR path (16 observations per second) into one observation per 5 minutes and added the GPS time points into this thinned time set. Notice that the thinning was done after the DR path was produced from the original 16Hz data set. The resulting Trip 1 data set had 2100 time points, among which 274 were GPS time points. The resulting Trip 2 data set had 2334 times points, among which 130 were GPS time points. It is noteworthy that our BM approach could easily be fitted to the original "big" data sets based on the 1Hz DR path (547803 time points for Trip 1 and 661249 for Trip 2). The results were reported in Liu et al. (2015) but omitted here as they are almost identical as in the following Table 4.4 or Figure 4.6 and 4.7. Individually modeling the two dimensions of the paths of the two animals yielded four data sets denoted as Trip 1 Northing (latitude), Trip 1 Easting (longitude), Trip 2 Northing, and Trip 2 Easting, respectively.

We considered different bias functions h in our BM approach as well as different random processes γ_1, γ_2 for the downscaling approach. To compare these different models as well as the two approaches, we conducted leave–5–out cross validation experiments (L5OCV). In the L5OCV, we removed 5 consecutive GPS observations at once when fitting our models and compared our model predictions based on the DR path to the original GPS observations. Such a cross validation scheme was designed to evaluate the predictive ability of our models when the time gaps between the GPS observations are of relatively large size. We were less concerned with the model's predictive ability for short gaps as there were natural constraints on the speed of our tracked animal. This means that when the time gaps were small, the animal's movements were confined in a small range and the performance of all methods were similar, i.e., in leave–one–out cross valida-

tions (Liu et al., 2015). L5OCV in our real data sets created gaps that would be longer than roughly 90% of the gaps in the observed GPS time points as shown in Table 2.1 and thus provided us with a good way to evaluate the long term predictive performance. We used the root mean squared error (RMSE) as a measure for the prediction accuracy and also calculated the actual coverage percentage of the nominal 95% posterior credible intervals to examine whether the uncertainty in the combined path is calibrated properly.

In the rest of this section, the goodness of normal approximation credible interval (4.8) in the BM approach is first studied in Section 4.5.1. Next, we summarize the model selection results of the BM and downscaling models (Section 4.5.2 and 4.5.3 respectively). Then the cross–validation comparison with BM and downscaling approaches as well as linear interpolation, CRAWL, and conventional bias correction is described in Section 4.5.4. The corrected path and its credible intervals for both methods are included in Section 4.5.5. We further study the distance traveled by the animal in Section 4.5.6

4.5.1 Goodness of normal approximation credible intervals

The normal mixture scheme for numerical integration described in Appendix A.6 provides an almost "exact" way to calculate the point-wise credible intervals from $\eta(t)$, that is, we can numerically calculate quantiles of the normal mixture at each time point. However, calculating these exact credible intervals can be time consuming for a large number of time points. A simple solution is to approximate the exact normal mixture posterior distribution of η with a normal distribution with mean (A.18) and variance (A.19). In our case study, we find that using the normal approximation interval in (4.8) can provide almost identical credible intervals as the exact ones. Using the Trip 1 Northing direction as an example, we plotted the exact normal mixture density and normal approximation density for a few randomly picked time points in Figure 4.4. Similar plots were found for other data sets and time points, which are thus omitted.

From Figure 4.4, it is clear that the normal approximation density is indistinguishable from the exact density. There are two reasons for this. First, the "randomness" from $\boldsymbol{\phi}$ is relatively small comparing to the "randomness" of $\boldsymbol{\eta}$ con-

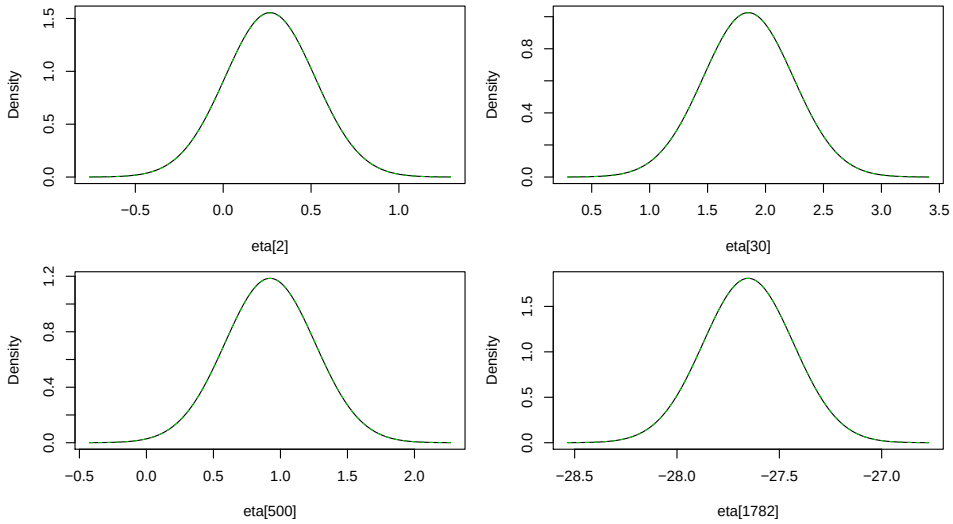


Figure 4.4: The exact credible posterior density (solid line) and normal approximation density (dotted line) for a few $\eta(t)$ in Trip 1 Northing direction for a northern fur seal.

ditioning on $\boldsymbol{\phi}$, i.e., the full Bayesian posterior $[\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}]$ is not very different from the empirical Bayesian posterior $[\boldsymbol{\eta}|\hat{\boldsymbol{\phi}},\mathbf{X},\mathbf{Y}]$, where $\hat{\boldsymbol{\phi}}$ is the posterior mode. As $[\boldsymbol{\eta}|\hat{\boldsymbol{\phi}},\mathbf{X},\mathbf{Y}]$ is normally distributed, it is not surprising to find that $[\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}]$ is nearly normal. Second, in our case study, $\boldsymbol{\phi}$ is well "estimated" from more than one hundred observations $\mathbf{X}_G,\mathbf{Y}_G$. The large sample size for two parameters guarantees that the posterior of $\boldsymbol{\phi}$ is also centered around the posterior mode and approximately normal. The full posterior $[\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}]$ can thus be viewed as a multivariate Gaussian resulting from a Gaussian distribution with Gaussian prior mean.

4.5.2 Model selection for BM

In our BM approach, we used the Brownian Bridge and Brownian Motion processes with different bias functions h in the DR path. Among many possible parameterizations of h, we investigated only the polynomials $h(t) = \sum_{i=1}^{Q} \beta_i t^{i-1}$ of order $Q = 1$ (constant) to $Q = 4$. The RMSE and actual coverage are shown in Table 4.1.

As seen in Table 4.1, the BM with different $h(\cdot)$ functions had very similar

RMSE, i.e., they differed little compared to the difference between the BM approach and linear interpolation, as in Table 4.4. The actual coverages of the credible intervals were reasonably close to the nominal 95% among the different Q's. This indicates that increasing the complexity in $h(\cdot)$ had little impact on the performance of our BM approach in our data sets. There was an exception for the Trip 2 Northing direction where increasing Q to 3, 4 reduced the CVRMSE, but it came at the cost of lowering coverage percentages for the credible intervals. This observation led us to choose the simple BM approach with $Q = 1$ for further comparisons.

Table 4.1: RMSEs and actual coverage percentages of 95% credible intervals (in gray background) in L5OCV comparisons for different bias correction term $h(t) = \sum_{i=1}^{Q} \beta_i t^{i-1}$ with $Q = 1, 2, 3, 4$ in the BM approach.

	Q=1		Q=2		Q=3		Q=4	
Trip 1 Northing	0.80	94.9	0.80	95.2	0.80	95.6	0.80	95.6
Trip 1 Easting	0.75	97.8	0.75	98.2	0.76	97.8	0.76	97.8
Trip 2 Northing	3.06	93.0	3.06	93.0	2.73	92.2	2.83	89.1
Trip 2 Easting	2.62	96.9	2.60	96.1	2.52	93.8	2.53	93.8

Perhaps our findings of little difference among the different BM models should not seem surprising, as we marginalized over the variance parameters σ_H^2, σ_D^2 and $\boldsymbol{\beta}$ (parameters in h) when evaluating the posterior $[\boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}]$. That marginalization naturally helps reduce reliance on a good mean model to correct the bias in the DR path. Therefore, we chose not to pursue that investigation for more sophisticated parameterization of h.

4.5.3 Model selection for downscaling

For the downscaling model (4.16), we considered random walks of order 1 and 2 (RW1, RW2) and autoregressive processes of order 1, 2, 3 (AR1, AR2, AR3) for both γ_1 and γ_2, leading to 25 models in total. They are denoted in "XXX–YYY" form, (i.e., RW1–AR2 denotes the downscaling model with γ_1 as a random walk of order 1 and γ_2 as an autoregressive process of order 2). Not every model fit our data sets well and INLA failed to converge due to numerical issues when fitting certain models. It was not feasible with our computational resources to perform cross–validation for all of the 25 models. Instead, we screened these models using

the conventional deviance information criterion (Spiegelhalter et al., 2002, DIC), which can also be calculated by the INLA package. We found that the DIC was minimized by some simple models for our data sets and have listed the DIC values for them together with the simplest RW1–RW1 model in Table 4.2.

Table 4.2: Selected DIC values for the downscaling models. "NA" means that INLA crashed when fitting this model. The AR1–AR2 model minimizes the DIC for Trip 1 Northing, AR2–RW1 for Trip 1 Easting and Trip 2 Easting, and RW2–RW1 for Trip 2 Northing. The RW1–RW1 model is included as a benchmark.

	RW1–RW1	RW2–RW1	AR1–AR2	AR2–RW1
Trip 1 Northing	-1608.41	-51.87	-1641.93	-1542.40
Trip 1 Easting	-487.04	-367.29	NA	-1683.86
Trip 2 Northing	-243.34	-853.31	-729.44	-827.04
Trip 2 Easting	-774.80	-91.57	-741.25	-829.27

None of the models involving AR3 components were selected by the DIC criterion, indicating that the dependence in the two random process were of short memory, i.e., they only depended on the previous two time points. Also, it is clear from Table 4.2 that none of these models dominated in all four data sets. The RW2–RW1 model achieved the smallest DIC in the two Easting data sets, but it did not fit well in the Trip 1 Northing. However, the simplest RW1–RW1 model achieved reasonable fit in all of the four data sets and thus was included in the following comparisons.

As pointed out by Spiegelhalter et al. (2014), DIC represents how the model fits the observed data with certain penalty on the model complexity and is not an ideal criterion for the predictive power of the models, which was our key objective in reconstructing the animal's path. We further compared the four downscaling models via a leave–5–out cross validation and summarized the results in Table 4.3. Among the four downscaling models we considered, AR2–RW1 and AR1–AR2 were the first to be ruled out as they had large prediction errors in the Trip 2 Northing and Easting data sets. Close examination of the cross–validation results found they completely failed to correct the DR path for a certain period of the trip and resulted in errors as large as 100km. In addition, the credible intervals for AR2–RW1 and

AR1–AR2 failed to achieve the nominal coverage percentage for Trip 1 Northing or Trip 2 Easting. As for the comparison between the RW2–RW1 and RW1–RW1 models, RW2–RW1 had slightly smaller RMSEs in the two Easting data sets while RW1–RW1 had smaller RMSEs in the two Northing data sets. Yet these two models, in general, achieved similar performances in terms of the RMSE and actual coverage percentage. We therefore chose the simpler RW1–RW1 model for further comparisons.

Table 4.3: RMSEs and actual coverage percentages of nominally 95% credible intervals (in gray background) for the L5OCV comparisons of the different downscaling models with different processes.

		Downscaling with different γ_1, γ_2			
		RW1–RW1	RW2–RW1	AR1–AR2	AR2–RW1
Trip 1 Northing	RMSE (km)	0.95	0.83	1.06	2.73
	Coverage (%)	95.6	91.5	94.9	86.0
Trip 1 Easting	RMSE (km)	0.75	0.93	0.71	0.80
	Coverage (%)	98.9	95.2	96.3	96.6
Trip 2 Northing	RMSE (km)	2.59	1.61	3.56	5.26
	Coverage (%)	93.0	98.4	92.9	91.1
Trip 2 Easting	RMSE (km)	2.56	4.08	18.27	18.32
	Coverage (%)	98.4	93.8	85.2	96.1

4.5.4 Cross–validation comparison of the different approaches

We compared the selected BM model with $Q = 1$ (BM1 for short) and downscaling RW1–RW1 model (DS11 for short) with the competitors: linear interpolation, conventional bias correction, and CRAWL (Johnson et al., 2008). Linear interpolation provided the same mean curve as using the Brownian Bridge to interpolate the GPS observations. CRAWL interpolated the GPS observations via state space models, whose process model is an integrated Ornstein–Uhlenbeck process with drift term (but assuming the correlation and process noise parameters are the same in both dimensions). Linear interpolation and CRAWL were based on the GPS observations only. Conventional bias correction was previously described in (2.1). The CV–RMSE and coverage percentages of the credible intervals are summarized

in Table 4.4.

We first compared the two approaches only with the GPS observations (first two columns of Table 4.4) to the approaches that combined both the GPS observations and DR path (the last three columns of Table 4.4). We found that the latter approaches had better predictive performance in general, which demonstrates the great value of the DR path in providing fine scale details of the animal's movement. In the comparison of linear interpolation and CRAWL, the more complex CRAWL had a larger RMSE than linear interpolation in the two Easting data sets we considered, which indicates a poor fit for the CRAWL models. In addition, the coverage percentages of CRAWL were lower than the nominal level in the two data sets from Trip 2.

We also noticed that the conventional approach had a larger RMSE than linear interpolation with the Trip 1 Northing data set, which shows that the conventional approach failed to properly use the high resolution DR path. The same issue was not found with our BM and downscaling approaches, both of which achieved a smaller RMSE than the other three approaches uniformly in all the four data sets we considered. Also the reduction in the RMSE achieved by our BM or downscaling approaches was larger than the differences between the BM or downscaling approaches with different model components as shown in Tables 4.1 and 4.3.

In the comparison between the selected BM and downscaling models (last two columns of Table 4.4), the BM1 had a smaller RMSE for Trip 1 Northing while the DS11 had a smaller RMSE for the Trip 2 Northing. They achieved similar RMSEs in the two Easting data sets. They also had similar coverage percentages across all four data sets. Thus, we conclude that the BM and downscaling approaches have similar performance.

To further explain the results in Table 4.4, we plotted the corrected path for all five approaches considered above and zoomed in on the time period 12:00—24:00 for 2009-07-23 in Trip 1 to better illustrate their differences (Figure 4.5). Similar plots were obtained by zooming into other periods and thus omitted. From this plot, we see that the conventional corrected path went through the GPS observations directly as did the linear interpolation; but it inferred dramatic changes between the GPS observations. The reconstructed path from BM and downscaling aligned well with the GPS observations, while retaining detail from the conventional bias cor-

Table 4.4: RMSEs and actual coverage percentages of nominally 95% credible intervals (in gray background) for the L5OCV comparisons of all the approaches. Two approaches that only use GPS data: Linear interpolation (Linear), and Correlated Random Walk (CRAWL) with drift term. Three approaches use both GPS and DR path: conventional bias correction of DR (Convention), Bayesian Melding with $Q = 1$ (BM1), and Downscaling with RW1–RW1 model (DS11).

		Linear	CRAWL	Convention	BM1	DS11
Trip 1 Northing	RMSE (km)	1.16	1.12	1.25	0.80	0.95
	Coverage (%)		94.1		94.9	95.6
Trip 1 Easting	RMSE (km)	1.13	1.28	1.04	0.75	0.75
	Coverage (%)		93.8		97.8	98.9
Trip 2 Northing	RMSE (km)	4.44	3.70	3.33	3.06	2.59
	Coverage (%)		88.3		93.0	93.0
Trip 2 Easting	RMSE (km)	3.84	3.98	2.67	2.62	2.56
	Coverage (%)		90.6		96.9	98.4

rection. As discussed in Section 4.2.2, the BM corrected path shrinks the conventional bias correction toward that of linear interpolation. This shrinkage removes the noise in the DR path in a statistical way and gives the BM an advantage over the conventional approach. From this plot, we also found that the corrected path from BM and downscaling were very close.

However, the path from CRAWL showed some unrealistic trends, like the upswing before 19:00, while the DR path indicated that the animal was moving in the opposite direction. These unrealistic trends likely resulted from the model assumptions in CRAWL, which may not fit the data well. These unrealistic trends and the resulting poor performance of CRAWL in cross–validations seem to have derived from the lack of fine detail provided by the DRA.

4.5.5 Combined paths and their uncertainty from BM and downscaling

Figure 4.5 only covered a small period of our data. We applied our proposed BM and downscaling approaches to the four data sets and found they all successfully corrected the bias of the DR path and properly quantified the uncertainty. There-

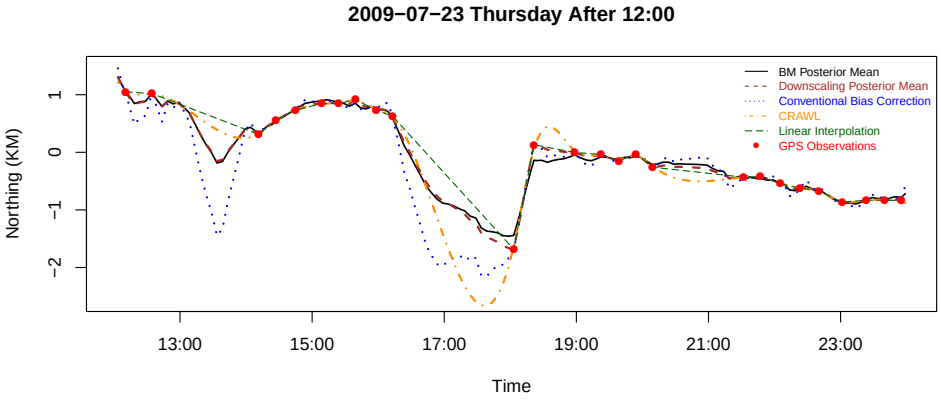


Figure 4.5: The reconstructed path for a northern fur seal in a selected period from all the five approaches considered in our study, Bayesian Melding (BM1), Downscaling (DS11), Conventional bias correction, CRAWL and linear interpolation. The dots are the GPS observations.

fore, we only present the plots for the corrected path for the Trip 1 Northing data set. Similar plots and analysis were found in the other three data sets and are thus omitted.

In Figure 4.6, we show the corrected path from the BM (solid curve) and downscaling (dotted curve) approaches, which are the posterior mean of $\eta(t)$ when BM was used, and $Y(t)$ when the downscaling approach was used. The point-wise 95% credible intervals (gray solid curve for BM and purple dotted curve for downscaling) were included to represent the uncertainty around the corrected path. We also included the original data—GPS observations (round points) and the DR path (dashed curve), from which it is clear that the bias of the DR path grew dramatically over time and reached 100km at the end of this trip. The location estimate in the DR path was not very useful in predicting the animal's location, but the fluctuations in the DR path matched the fluctuations of the GPS observations, meaning that the DR path has useful high–frequency information that can be further exploited to fill in the gaps between GPS observations. This was successfully achieved by our proposed BM and downscaling approaches, as the corrected path from both approaches lay close to the GPS observations.

For the scale of Figure 4.6, the corrected path and CIs from BM and down-

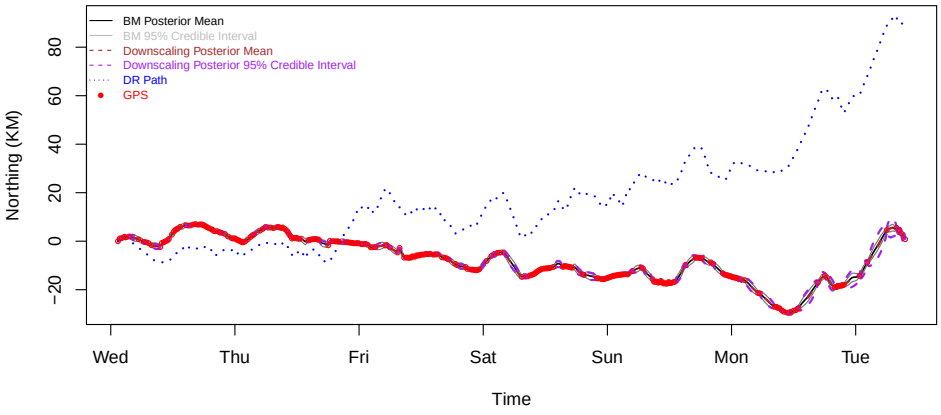


Figure 4.6: The Bayesian Melding and downscaling results for a northern fur seal undertaking Trip 1 Northing from Bogoslof Island, Alaska: Points are the GPS observations and the dotted curve is the DR path. The solid curve is the posterior mean of $\eta(t)$ in the case of BM, whose 95% credible intervals are shown by the gray solid curve. The dotted curve is the posterior mean of $Y(t)$ in the downscaling approach, whose 95% credible intervals are shown by the gray dotted curve.

scaling were almost indistinguishable. To show how our BM and downscaling approaches worked in fine scale, we zoomed into the Day 2 (2009–07–23) and Day 6 (2009–07–27) part of this trip. We can confirm from Figure 4.7 that the corrected paths from the two approaches were similar as the curves of the posterior means nearly overlaid each other. As well, the CIs from both approaches displayed a clear "bridge" structure, that is, they were narrower at the GPS time points and wider in between the GPS observations. This is plausible because we have direct and accurate observations at the GPS time points and less accurate information when the GPS locations were not available. The error grows as the track moves away from the GPS observations and decreases near the next GPS observation. Also, the longer the gap between the GPS time points, the larger the error, and thus, the wider the credible intervals, which is seen by comparing Days 2 and 6. On Day 6, fewer GPS observations were available and the gaps were longer, which resulted in overall wider credible intervals.

Another interesting finding seen in Figure 4.7 is that the CI from the downscal-

ing approach was narrower than those from BM on Day 2 while they were wider on Day 6. This was caused by the two RW1 components in the downscaling model, as their variance was growing with time (Figure 4.8). On the other hand, the posterior SD for the BM approach was more stable because it was more constrained with the Brownian Bridge structure.

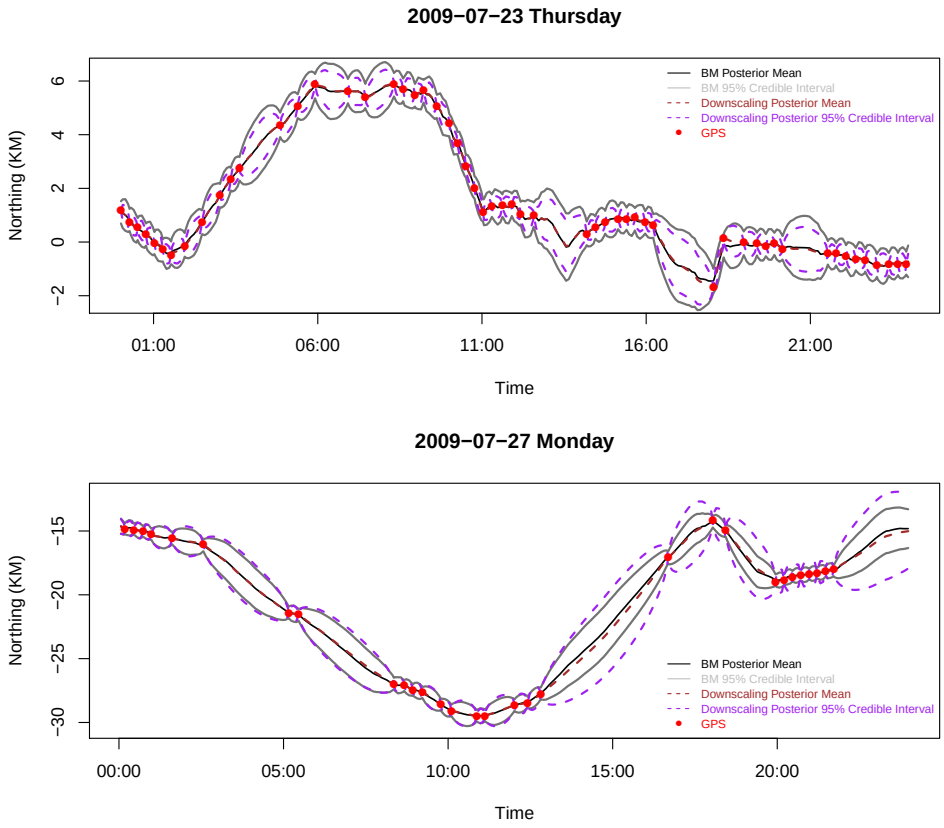


Figure 4.7: Zoomed Bayesian Melding and downscaling results for a northern fur seal undertaking Trip 1 Northing on 2009–07–23 and 2009–07–27: Red points are the GPS observations. The solid curve is the posterior mean of $\eta(t)$ in BM, whose 95% credible intervals are shown by the gray solid curve. The dotted curve is the posterior mean of $Y(t)$ in downscaling, whose 95% credible intervals are shown by the gray dotted curve.

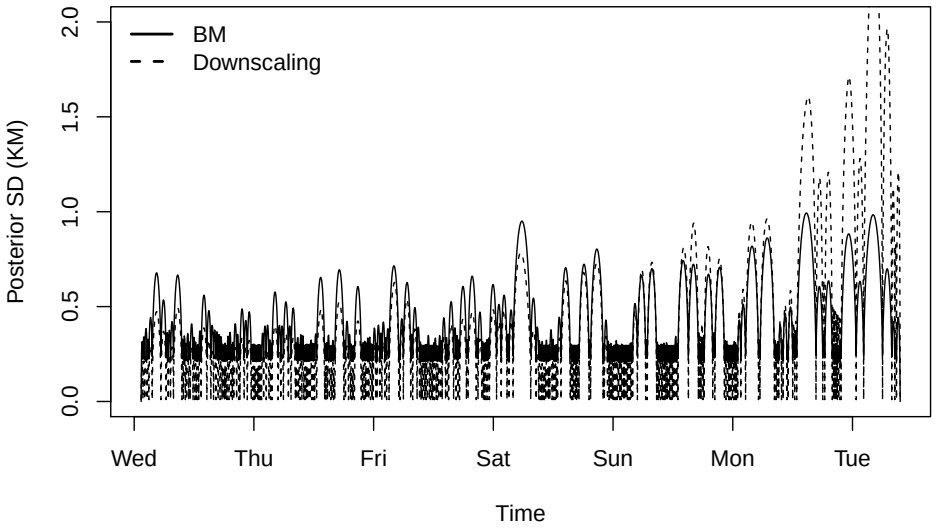


Figure 4.8: Posterior standard deviation (SD) from Bayesian Melding and downscaling results for a northern fur seal undertaking Trip 1 Northing. The solid curve is from BM while dotted curve is from downscaling.

4.5.6 Paths in both dimensions and the distance traveled by the animal

We plot the corrected paths from the BM approach in both dimensions for these two trips in Figure 4.9 and omit the similar paths from the downscaling approach, because the correct paths from downscaling were almost indistinguishable from those the BM on the scale of the whole trip (seen in Figure 4.6). From Figure 4.9, it is easy to find that the corrected paths from our approaches not only connect the GPS observations, but also reserve the tortuosity exhibited by the DR path.

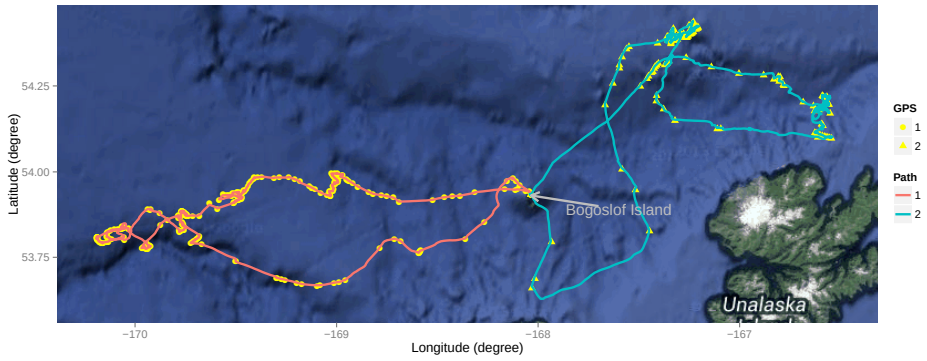
With these corrected paths from our BM approach, we more accurately calculated the distance traveled by the fur seals during their foraging trip (Table 4.5). Distances calculated using our approach were 40% greater for Trip 1 and 49% greater for Trip 2 than those calculated by linear interpolation of the GPS observations. This, once again, demonstrates that the corrected DR path is needed to fill in the gaps between the GPS observations. Calculating the distance traveled by the animal with the conventional bias correction method revealed it to be twice the distance calculated based on GPS observations, which further illustrates that our

BM is a compromise between linear interpolation and conventional DR correction.

Table 4.5: Total distance (km) traveled in the two fur seal foraging trips calculated via different methods

Trip	Linear Interpolation	BM	Conventional
1	418.25	585.95	815.36
2	443.30	662.91	1023.88

Figure 4.9: GPS observations of the two trips and the corrected paths from our BM approach. The GPS measured latitude and longitude of the two trips of the northern fur seal began and ended at Bogoslof Island in the summer of 2009. The GPS observations in Trip 1 are indicated by the round points and those for Trip 2 are indicated by the triangular points. The pink curve is the corrected path from our BM model for Trip 1 and the blue curve is for Trip 2.



4.6 Discussion

We present a Bayesian Melding approach and a downscaling approach to combine sparse but accurate GPS observations with high resolution but biased DR paths for the tracking of marine mammals. The posterior mean from our BM and down-

scaling approaches both offer an accurate and high resolution path for the tracked animals and the posterior credible intervals provide a reasonable statement of the uncertainty in our inferences. The good predictive performance of our approaches is also supported by our simulation studies. Moreover, neither of our methods requires computationally expensive MCMC methods for computation. Our BM approach exploits the conditional independence property of the Brownian Bridge and Brownian Motion to dramatically reduce the heavy computational burden involved in dealing with large data sets. The downscaling approach is fitted via the computationally efficient INLA approach. The quality of the likelihood approximation in BM and correctness of downscaling were confirmed in our simulation study.

We performed cross–validation studies to compare different models in these two approaches and found that the predictive performance of the simplest approaches (i.e., BM with a constant bias term and downscaling with both two random effects being first order random walks) were as good as or even better than those of the more complex models, according to our empirical assessments. This finding is partially explained by the fact that we marginalized over the model parameters in the posterior distribution for our tracked subjects. In the comparison between BM and downscaling, we could not find any noteworthy differences between these two approaches in their prediction accuracy and actual coverage percentage of their credible intervals. However, our implementation of BM is better because it is more scalable to big data sets with more than half a million time points on a regular computer (Liu et al., 2015), while the downscaling approach fitted by INLA can only work with thinned data sets on the same computer. Also, we can build BM on a process that reflects the nature of our tracked subject, which is discussed in Chapter 6.

McClintock et al. (2014) have shown many disadvantages of using a discrete time formulation when working with satellite data, which may lead one to question the discrete time formulation we used. Yet, animal movement, (e.g., the fur seals we tracked) is ultimately powered by its body movement (e.g., the stroking of flippers), and there is a maximum frequency of body movement that an animal is capable of achieving. Therefore, the animal's movement is essentially discrete and can thus be sufficiently well described by a discrete process model with an

observation frequency of no less than twice the maximum frequency of the animal's body movement (Nyquist frequency, see e.g., Leis, 2011; Le and Zidek, 2006). The Nyquist frequency was taken into consideration in data collection, i.e., the original sampling frequency of the DR path was 16Hz in our northern fur seal data. The DR path thus captures the fine detail in an animal's movements and provides observations of the path in "continuous time" with respect to the animal. Our discrete time formulation in Bayesian Melding or downscaling is thus backed up by these "continuous time" observations, which are free of the shortcomings of discrete time models for satellite data. In addition, the Bayesian Melding with Brownian Bridge and Brownian Motion have both processes defined in continuous time, which can be easily modified for a continuous time model.

One concern about our Brownian Bridge prior in the Bayesian Melding approach is that the GPS observations (Figure 4.1) appear to be much smoother than a Brownian Bridge process. Yet, it is important to recognize that the Brownian Bridge is only a prior distribution for the animal's path. The corrected path, as seen in Figure 4.7, retains the smoothness and does not become tortuous as a simulated Brownian Bridge, because the corrected path is the posterior given both the GPS and DR paths. The smoothness of the DR path is preserved in the posterior. However, our BM approach can undoubtedly be further improved by replacing the Brownian Bridge process prior with some other processes, which is the focus of the following chapters.

Chapter 5

Conditional Heterogeneous Gaussian Process

In this chapter, we propose a conditionally heterogeneous Gaussian process (CHGP) for the modeling of inhomogeneous spatio–temporal data. Here "inhomogeneous" means that the process have non-stationary covariance function with changing variance or correlation parameters. For GP defined on the time domain, "inhomogeneous" indicates the process has non-constant variance/correlation parameters. In a CHGP, a Gaussian process (GP) is specified first on a set of knots. Conditioning on these knots, the sub-intervals separated by the knots follow GP's of the same family but with possibly different parameter values. The knots and sub-intervals should enable us to handle efficiently large data sets and the "conditionally heterogeneous" parameters can capture the non-stationary dependence structure of the data. It is also possible to learn the hidden behavior from based on different conditional parameters. We first study a special case of the CHGP, the conditionally heterogeneous Brownian Bridge (CHBB) as an initial investigation. Many metrics and algorithms are considered for the model selection in the CHBB modeling, but none of them succeeds in choosing a reasonable knot structure automatically. This chapter documents our attempts and discusses the reasons why this is an unsuccessful idea.

The rest of this chapter is organized as follows. Section 5.1 introduces the CHGP and CHBB processes and their relationship to some other recent devel-

opments in GP modeling. Section 5.2 introduces and discusses our metrics and algorithms for model selection in the CHBB modeling. Our reflections on these efforts are discussed in Section 5.3 .

5.1 Definition of CHGP and CHBB

As in Section 3.1, we find that most of the recent developments in spatio–temporal modeling with the GP are focused on either the non-stationarity issue or the big data issue and few methods are good at dealing with both issues. Our proposal of the conditionally heterogeneous GP (CHGP) is designed to simultaneously handle the non-stationarity and big data issues. Similarly as in the NNGP modeling (Datta et al., 2016), we begin with a set of knots in $\mathbb{R}^d$, $\mathscr{S} = \{\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_K\}$ and define a multivariate Gaussian random vector,

$$\boldsymbol{\eta}(\mathscr{S}) \sim \mathrm{N}(\mathbf{f}(\mathscr{S}), \mathbf{C}(\mathscr{S}, \mathscr{S}; \boldsymbol{\theta}_0)),$$

where $\mathbf{f}(\mathscr{S})$ is the mean vector and $\mathbf{C}(\mathscr{S}, \mathscr{S}; \boldsymbol{\theta}_0)$ is the covariance matrix. In addition to the knots, we also respectively observe or aim to predict the process at locations $\mathscr{T} = \{\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_K,\ \mathbf{s}_{K+1}, \mathbf{s}_{K+2}, \ldots, \mathbf{s}_N\}$. We assume that the knots $\mathscr{S} \subset \mathscr{T}$ for simplicity in discussion, but the model will work as well without this assumption. We divide the non-knot locations $\mathscr{T} \setminus \mathscr{S} = \{\mathbf{s}_{K+1}, \ldots, \mathbf{s}_N\}$ into J disjoint blocks $\mathscr{T}_j, j = 1, 2, \ldots, J$, such that

$$\begin{aligned} &\mathscr{T}_j \cap \mathscr{T}_{j'} = \emptyset, && \text{for } j \neq j', \\ &\bigcup_{j=1}^{J} \mathscr{T}_j = \mathscr{T} \setminus \mathscr{S}. \end{aligned}$$

Conditioning on $\boldsymbol{\eta}(\mathscr{S})$ and $\boldsymbol{\theta}_0$, $\boldsymbol{\eta}(\mathscr{T}_j)$ is assumed to be another GP

$$\boldsymbol{\eta}(\mathscr{T}_j) | \boldsymbol{\eta}(\mathscr{S}) \sim \mathrm{N}(\mathbf{F}_j \boldsymbol{\eta}(\mathscr{S}_j), \mathbf{C}_j^*(\mathscr{T}_j, \mathscr{T}_j; \boldsymbol{\theta}_{1,j}, \boldsymbol{\theta}_0)),$$

where $\boldsymbol{\theta}_{1,j}$ denotes the additional parameters in the conditional covariance matrix $\mathbf{C}_j^*$. For example, $\mathbf{F}_j$ and $\mathbf{C}_j^*$ can be specified as in the Gaussian Predictive Pro-

cess (GPP, Banerjee et al., 2008):

$$\mathbf{F}_j = \mathbf{C}(\mathscr{T}_j, \mathscr{S}; \boldsymbol{\theta}_0)\mathbf{C}^{-1}(\mathscr{S}, \mathscr{S}; \boldsymbol{\theta}_0) \tag{5.1}$$

$$\mathbf{C}_j^* = \mathbf{C}(\mathscr{T}_j, \mathscr{T}_j; \boldsymbol{\theta}_{1,j}) - \mathbf{C}(\mathscr{T}_j, \mathscr{S}; \boldsymbol{\theta}_0)\mathbf{C}^{-1}(\mathscr{S}, \mathscr{S}; \boldsymbol{\theta}_0)\mathbf{C}(\mathscr{S}, \mathscr{T}_j.\boldsymbol{\theta}_0). \tag{5.2}$$

In this case, $\boldsymbol{\theta}_{1,j}$ are the parameters in the marginal covariance matrix for $\boldsymbol{\eta}(\mathscr{S}_j)$. We can also consider other parameterizations for $\mathbf{F}_j$ and $\mathbf{C}_j^*$, such as the radial basis function[1] from the multi-resolution GP (MRGP, Nychka et al., 2015).

It is easy to verify that the CHGP reduces to a GP when $\mathscr{S} = \emptyset$, or $\mathscr{S} = \mathscr{T}$, or $\boldsymbol{\theta}_{1,j} = \boldsymbol{\theta}_0, \forall j$ in expression (5.2). If $\mathbf{F}_k, \mathbf{C}_k^*$ are specified in the GPP manner as above, we can choose $\mathscr{T}_k$ to be the subset whose elements have the same m-nearest neighbors in $\mathscr{S}$. If $\mathbf{F}_k, \mathbf{C}_k^*$ are specified via radial basis functions as in MRGP, $\mathscr{T}_k$ can be chosen to be a subset whose elements have non-zero correlation with the same set of elements in $\mathscr{S}$. There can be different ways of choosing the $\mathscr{T}_k$'s and it should be chosen to accommodate the non-stationary dependence structure. In addition to the one layer of knots as specified above, we can consider multiple layers of knots, i.e., specify another knot set inside $\mathscr{T}_j$, similarly as in Katzfuss (2016).

A key difference in the definition of the CHGP compared with the NNGP (Datta et al., 2016) and the MRGP (Nychka et al., 2015) is that the CHGP allows the parameters $\boldsymbol{\theta}$ to vary among the conditional sets $\mathscr{T}_j$, which is designed for inhomogeneous modeling. Also, instead of specifying a fixed set of knots or the nearest neighbors as in MRGP or NNGP, we attempt to compare the model with different knots and let the data choose an "optimal" set of knots. Our CHGP differs from the independent block (Stein et al., 2004) or the composite likelihood approach (Eidsvik and Shaby, 2014) due to the fact that the correlation between blocks in CHGP is still carried by the knot set.

A special case of the CHGP defined for time interval $\mathscr{T} = [0, T]$ is the conditionally heterogeneous Brownian Bridge (CHBB) process. At the knots $\mathscr{S} = \{s_0 = 0 < s_1 < s_2 \ldots s_K = T\}$, the random vector has the same joint distribution a

[1] Radial basis function is a basis function that only depends on the distance.

Brownian Bridge with variance parameter σ_0^2,

$$\boldsymbol{\eta}(\mathscr{S}) = \boldsymbol{\eta}(\mathbf{s}_{0:K}) \sim \mathrm{BB}\,(0,0,s_0,s_K,\sigma_0^2), \tag{5.3}$$

where $\mathrm{BB}\,(A,B,t,t',\sigma^2)$ is defined the same as in (4.3). The start and end points $\eta_0 = 0, \eta_T = 0$ are fixed for the definition of the BB and the knot set $\mathscr{S}$ always contains $0,T$. The knot set $\mathscr{S}$ separates $\mathscr{T}$ into K disjoint intervals (periods) $\mathscr{T}_k = (s_{k-1}, s_k), k = 1,\ldots,K$. The process in each period is a Brownian Bridge process with (conditional) variance parameter $\sigma_{1,k}^2$ conditioning on $\eta(s_{k-1}), \eta(s_k)$:

$$\boldsymbol{\eta}(\mathscr{T}_k)|\eta(s_{k-1}), \eta(s_k) \sim \mathrm{BB}\,(\eta(s_{k-1}), \eta(s_k), s_{k-1}, s_k, \sigma_{1,k}^2). \tag{5.4}$$

In addition, we assume conditional independence of any $\eta(t)$ and $\eta(t')$ when t and t' is separated by a knot s_k,

$$\eta(t) \perp \eta(t')|\eta(s_k), \qquad\qquad \forall, t < s_k < t'.$$

Denote the set of all the variance parameters by $\boldsymbol{\sigma}^2 = \{\sigma_0^2, \sigma_{1,1}^2, \ldots, \sigma_{1,K}^2\}$. Then the joint distribution of CHBB $[\boldsymbol{\eta}|\mathscr{S}, \boldsymbol{\sigma}^2]$ is

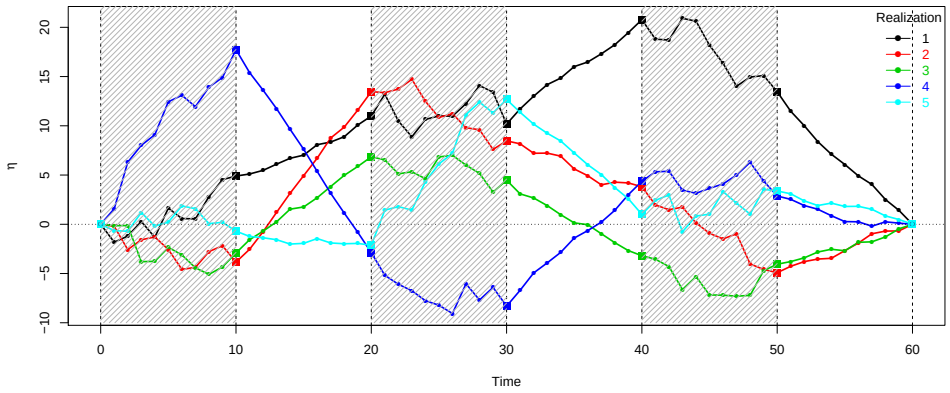
$$[\boldsymbol{\eta}|\mathscr{S}, \boldsymbol{\sigma}^2] = [\boldsymbol{\eta}(\mathscr{S})|\sigma_0^2] \prod_{k=1}^{K} [\boldsymbol{\eta}(\mathscr{T}_k)|\eta(s_{k-1}), \eta(s_k), \sigma_{1,k}^2]. \tag{5.5}$$

The joint distribution of CHBB can be viewed as the distribution of a Gaussian process with changing variance parameters. Clearly, a CHBB reduces to a Brownian Bridge process when $\sigma_0^2 = \sigma_{1,k}^2, \forall k$ or $\mathscr{S} = \mathscr{T}$ or $\mathscr{S} = \{0,T\}$. Notice that $\sigma_{1,k}^2$ can be different from σ_0^2, which allows this process to have inhomogeneous features and adapt to the local data. For modeling of an animal's track, σ_0^2 in CHBB can reflect the overall range of the foraging trip and the different $\sigma_{1,k}^2$ can reflect the animal's hidden states, e.g., foraging, traveling or sleeping during this trip.

As an illustration, Figure 5.1 plots five simulated realizations of a CHBB observed at $\mathscr{T} = \{0,1,\ldots,60\}$ with knot set $\mathscr{S} = \{0,10,\ldots,60\}$. The start and end points of these 5 realizations are fixed at 0 and the variance parameters are set to be $\sigma_0^2 = 10$, $\sigma_{1,1}^2 = \sigma_{1,3}^2 = \sigma_{1,5}^2 = 2$ and $\sigma_{1,2}^2 = \sigma_{1,4}^2 = \sigma_{1,6}^2 = 0.1$. From Figure 5.1, it is easy to see that the overall range of these CHBB's is mainly decided by σ_0^2. Also,

in the shaded periods, the non-knot points are far away from the lines connecting the knots because of a large $\sigma^2_{1,k}$. On the other hand, in the non-shaded periods, the non-knot points stay relatively close to the lines connecting the knots. For the marine mammal tracking application, the shaded periods may represent the animal's foraging behavior, where the animal searches its surroundings for food while the non-shaded periods may present animal's traveling behavior, where it swims towards a certain destination, such as its home.

Figure 5.1: Five simulated realizations of a CHBB with knot set $\mathscr{S} = \{0, 10, 20, \ldots, 60\}$. The start and end points of this CHBB are fixed at zero. The variance parameters are chosen to be $\sigma^2_0 = 10$, $\sigma^2_{1,1} = \sigma^2_{1,3} = \sigma^2_{1,5} = 2$ (shaded areas) and $\sigma^2_{1,2} = \sigma^2_{1,4} = \sigma^2_{1,6} = 0.1$. The vertical dashed lines indicate the knots $\mathscr{S}$.



The definition of CHBB is different from the dynamic Brownian Bridge process in Gurarie et al. (2009) as well as Kranstauber et al. (2012, 2014), which assumes that the variance parameter is changing with time. They choose to estimate the variance parameter at time t by averaging over the estimates from all the moving windows that contain t. This produces a sequence of highly correlated variance parameters and it is unclear to us how to choose the window size. If the window covers observations from two behavior status of the animal, it is difficult to explain what the variance parameter represents.

5.2 Attempts at model selection for the CHBB

An important feature of the CHBB and CHGP is that we allow the data to decide the knots and the conditional parameters. The conditional parameters can be estimated given the knots and thus the key issue is how to select the knots $\mathscr{S}$, which involves two issues: how to enlist the candidate $\mathscr{S}$'s and how to measure their differences. For the first issue, we consider sequentially adding knots to an empty knot set, or deleting knots from a full knot set, because $\mathscr{S} = \{0, T\}$ or $\mathscr{T}$ are the same model. We find empirically that the search direction has little impact on the selected knots once the metric for $\mathscr{S}$ is decided. As a sparse knot set is preferred, sequentially adding knots is more efficient and used in our following discussion.

5.2.1 Metrics for model comparison

To measure the differences between the CHBB models with different knots, we have considered the following metrics: the likelihood ratio (LR) statistic, Kullback–Leibler (KL) divergence and three Bayes factors: original, intrinsic, and fractional. We use M to denote the CHBB model, including the knots $\mathscr{S} \subseteq \mathscr{T}$ and the variance parameters $\boldsymbol{\sigma}^2 = \{\sigma_0^2, \sigma_{1,1}^2, \ldots, \sigma_{1,K}^2\}$. The bracketed subscript on $M_{(i)}$ indexes the different models.

The LR statistic and KL divergence are two frequentist metrics for model comparisons. The LR statistic D is twice the difference between the likelihood of two nested models $M_{(0)}, M_{(1)}$

$$D = -2\bigg(\log([\mathbf{y}|\mathscr{S}_{(0)}, \hat{\boldsymbol{\sigma}}^2_{(0)}]) - \log([\mathbf{y}|\mathscr{S}_{(1)}, \hat{\boldsymbol{\sigma}}^2_{(1)}]) \bigg), \tag{5.6}$$

where $[\mathbf{y}|\mathscr{S}_{(i)}, \boldsymbol{\sigma}^2_{(i)}], i = 0, 1$ is the joint probability of $\mathbf{y}$ as in (5.5) and $\hat{\boldsymbol{\sigma}}^2_{(i)}, i = 0, 1$ are the maximum likelihood estimates of $\boldsymbol{\sigma}^2$given $\mathscr{S}$ and $\mathbf{y}$. Model $M_{(0)}$ is nested in $M_{(1)}$ when $\mathscr{S}_{(0)} \subset \mathscr{S}_{(1)}$.

The KL divergence measures the discrepancy between $M_{(1)}$ and $M_{(0)}$. As with the LR statistic, we evaluate the divergence at the MLE of the parameter estimates,

$$\mathrm{KL}(M_{(0)}||M_{(1)}) = \int [\mathbf{y}|\mathscr{S}_{(0)}, \hat{\boldsymbol{\sigma}}^2_{(0)}] \log \left(\frac{[\mathbf{y}|\mathscr{S}_{(0)}, \hat{\boldsymbol{\sigma}}^2_{(0)}]}{[\mathbf{y}|\mathscr{S}_{(1)}, \hat{\boldsymbol{\sigma}}^2_{(1)}]} \right) d\mathbf{y}. \tag{5.7}$$

A shortcoming of the LR statistic and KL divergence is that they depend on the MLE of the parameters and do not account for the uncertainty in the parameters. This shortcoming can be overcome by using the Bayes factor (BF), which compares two models by the marginal probability of the data y,

$$[\mathbf{y}|M_{(i)}] = \int [\mathbf{y}|\mathscr{S}_{(i)}, \boldsymbol{\sigma}^2_{(i)}][\boldsymbol{\sigma}^2_{(i)}] d\boldsymbol{\sigma}^2_{(i)}, \qquad i = 0, 1, \tag{5.8}$$

where $[\boldsymbol{\sigma}^2_{(i)}]$ is the prior for variance parameter $\boldsymbol{\sigma}^2_{(i)}$. The Bayes factor is defined as the ratio of the two marginal distributions

$$\mathrm{BF}(M_{(1)}, M_{(0)}) = \frac{[\mathbf{y}|M_{(1)}]}{[\mathbf{y}|M_{(0)}]}. \tag{5.9}$$

Notice that a proper prior for the variance parameter is needed in the calculation of the marginal distribution and BF (O'Hagan, 1995). In the absence of the prior knowledge needed for an informative prior, the improper (uninformative) prior $[\sigma^2] \propto 1$ or $[\log(\sigma^2)] \propto 1$ is a convenient choice, but these improper priors will lead to an ill–defined BF.

To illustrate this problem, we introduce a normalizing constant c and write the improper prior as $c[\sigma^2]$. Notice that c does not exist but we act as if $c[\sigma^2]$ is a proper prior. Consider $M_{(0)}$ as a BB without knots; it has one variance parameter ς^2_0. Its marginal density is

$$[\mathbf{y}|M_{(0)}] = c \int [\mathbf{y}|\varsigma^2_0][\varsigma^2_0] d\varsigma^2_0. \tag{5.10}$$

Meanwhile, $M_{(1)}$ is a CHBB with $K = 2$ (three knots) and thus it has three variance parameters $\sigma^2_0, \sigma^2_{1,1}, \sigma^2_{1,2}$. The marginal probability becomes

$$[\mathbf{y}|M_{(1)}] = c^3 \int [\mathbf{y}|\mathscr{S}_{(1)}, \sigma^2_0, \sigma^2_{1,1}, \sigma^2_{1,2}][\sigma^2_0][\sigma^2_{1,1}][\sigma^2_{1,2}] d\sigma^2_0 d\sigma^2_{1,1} d\sigma^2_{1,2}. \tag{5.11}$$

Dividing (5.11) by (5.10), the BF between $M_{(0)}$ and $M_{(1)}$ has a leading factor of c^2. This non-existent normalizing constant is part of the BF and thus demonstrates that it is inappropriate to use the BF with improper priors.

Two alternative forms of the BF have been proposed to construct the BF with

improper priors. The intrinsic BF (Berger and Pericchi, 1996, and references therein) first calculates the posterior with part of the data and then uses this posterior as the prior for the remaining data. A concern about the intrinsic BF is that users have to manually decide which observations to use in forming the prior. Such arbitrariness is avoided by the fractional BF (O'Hagan, 1995) which considers the following fractional marginal density

$$Q_b(M_{(i)}) = \frac{\int [\mathbf{y}|\mathscr{S}_{(i)}, \boldsymbol{\sigma}^2_{(i)}][\boldsymbol{\sigma}^2_{(i)}] d\boldsymbol{\sigma}^2_{(i)}}{\int [\mathbf{y}|\mathscr{S}_{(i)}, \boldsymbol{\sigma}^2_{(i)}]^b [\boldsymbol{\sigma}^2_{(i)}] d\boldsymbol{\sigma}^2_{(i)}},$$

where $b \in (0,1)$ is a small fraction that in effect decides how much data is used to "form the prior".

For the model selection of CHBB, we have considered all three forms of the BF. The original BF is applied with informative priors. Small fractions of $b = 1/T, 2/T, 3/T$, etc., are considered for the fractional BF. For the intrinsic BF, we use the posterior of knots $[\sigma_0^2|\mathbf{y}(\mathscr{S})]$ as the prior for $\mathscr{T} \setminus \mathscr{S}$. For all the metrics, LR, KL and the BFs, a larger value indicates a better fit of $M_{(1)}$. Yet, it is unclear how big an increase means a substantially better fit, except for the LR statistic, where the AIC or BIC can be used. We skip this issue by listing all candidate models that can be visited and studying the top models with biggest increases relative to the null model.

However, none of the above metrics offer satisfactory model selection results for both simulated and real data sets. They all tend to select two knots that are close to each other. This issue can be alleviated by a strongly informative prior in the original BF or a large b in fractional BF, but both have the undesired feature of sensitivity to the prior choices. For an individual data set, we may compare a few priors and choose one that yields a seemingly reasonable knot structure, i.e., matching our visual examination of the data. However, the prior comparison is redundant and it is easier to just pick the knots from the visual examination. Visual examination can also be challenging for big spatio–temporal data sets and prone to human bias.

5.2.2 Reason for their failure

To explain why all the metrics fail, we use the following example. Consider $\mathbf{y} = \{y_0, y_1, \ldots, y_T\}$ to be a realization of a CHBB at time set $\mathscr{T} = \{0, 1, 2, \ldots, T\}$. $M_{(0)}$ is the Brownian Bridge model,

$$M_{(0)} : \mathbf{y} \sim BB(y_0, y_T, 0, T, \varsigma^2).$$

The end points y_0, y_T are fixed. Model $M_1(s)$ has a single knot at time $s \in (1, T-1)$ (no empty periods),

$$M_{(1)} : \begin{cases} y_s | y_0, y_T \sim BB(y_0, y_T, 0, T, \sigma_0^2) \\ \mathbf{y}_{1:(s-1)} | y_0, y_s \sim BB(y_0, y_s, 0, s, \sigma_{1,1}^2) \\ \mathbf{y}_{(s+1):(T-1)} | y_s, y_T \sim BB(y_s, y_T, s, T, \sigma_{1,2}^2) \end{cases}$$

Under $M_{(0)}$, the twice negative log likelihood is

$$-2\log([\mathbf{y}|\varsigma^2]) = (T-1)\big(\log(2\pi) + \log(\varsigma^2)\big) + \log(|\mathbf{C}|) + (\mathbf{y}-\mathbf{f})^T \mathbf{C}^{-1} (\mathbf{y}-\mathbf{f}) \frac{1}{\varsigma^2},$$

where $\mathbf{C}$ is the covariance matrix of a standard (variance parameter as 1) Brownian Bridge on $\mathscr{T}$ and $\mathbf{f}$ is the mean of the Brownian Bridge as in (4.3). The MLE of ς^2 can be calculated as

$$\hat{\varsigma}^2 = \frac{1}{T-1} (\mathbf{y}-\mathbf{f})^T \mathbf{C}^{-1} (\mathbf{y}-\mathbf{f}). \tag{5.12}$$

Plugging the MLE $\hat{\varsigma}^2$ back into the log likelihood, we have,

$$-2\log([\mathbf{y}|\hat{\varsigma}^2]) = (T-1)\big(1 + \log(2\pi) + \log(\hat{\varsigma}^2)\big) + \log(|\mathbf{C}|). \tag{5.13}$$

Similarly, under $M_{(1)}$,

$$\begin{aligned}-2\log([\mathbf{y}|\hat{\sigma}_0^2,\hat{\sigma}_{1,1}^2,\hat{\sigma}_{1,2}^2]) =& \big(1+\log(2\pi)+\log(\tilde{\sigma}_0^2)\big)+\log(|\mathbf{C}_0|)\\ &+(s-1)\big(1+\log(2\pi)+\log(\hat{\sigma}_{1,1}^2)\big)+\log(|\mathbf{C}_{1,1}|)\\ &+(T-s-1)\big(1+\log(2\pi)+\log(\hat{\sigma}_{1,2}^2)\big)+\log(|\mathbf{C}_{1,2}|).\end{aligned} \tag{5.14}$$

The estimates of $\hat{\sigma}_0^2,\hat{\sigma}_{1,1}^2,\hat{\sigma}_{1,2}^2$ can be calculated following (5.12). The matrices $\mathbf{C}_0,\mathbf{C}_{1,1},\mathbf{C}_{1,2}$ are the covariances of a standard Brownian Bridge process for the corresponding time points.

Plugging (5.14) and (5.13) into (5.6), the LR statistic between $M_{(0)}$ and $M_{(1)}$ is

$$D=\log\left(\frac{\hat{\xi}^2}{\hat{\sigma}_0^2}\right)+(s-1)\log\left(\frac{\hat{\xi}^2}{\hat{\sigma}_{1,1}^2}\right)+(T-s-1)\log\left(\frac{\hat{\xi}^2}{\hat{\sigma}_{1,2}^2}\right). \tag{5.15}$$

A key simplification is the factorization of matrix determinant $|\mathbf{C}|=|\mathbf{C}_0||\mathbf{C}_{1,1}||\mathbf{C}_{1,2}|$. The proof is included in Appendix B.1.

With (5.15), we may further derive the joint distribution for Ds with different choices of knot s under the null model $M_{(0)}$, which is denoted as $D_s, s=2,3,\ldots,T-2$. The joint distribution $Pr\{D_2,D_3,\ldots,D_{T-2}\}$ can then be used to calculate $Pr\big\{\arg_t\max\{D_t\}=s\big\}$, namely, the probability that knot s is selected as the best $M_{(1)}$. However, those probabilities become very complicated to derive in a closed form and therefore we evaluate them by simulation. Figure 5.2 is the histogram of the selected s based on 1000 replicates of the BB with $T=100$. This histogram is U–shaped and the s's near the two end points, $s<5$ or $s>95$, are more likely to be selected than the middle s's. This illustrates the tendency that the model selection based on LR statistic is very likely to result in short periods in the CHBB.

The exact form of D in (5.15) provides useful heuristics into this problem with the LR statistic. When one of the variance estimates $\hat{\sigma}_0^2,\hat{\sigma}_{1,1}^2,\hat{\sigma}_{1,2}^2$ is close to zero, the metric becomes very large and two very close knots are more likely to produce a small variance estimate, because the sample size is smaller. For example, $\hat{\sigma}_{1,1}^2$ is the MLE of the variance parameter of the BB conditional on y_0 and y_s, which

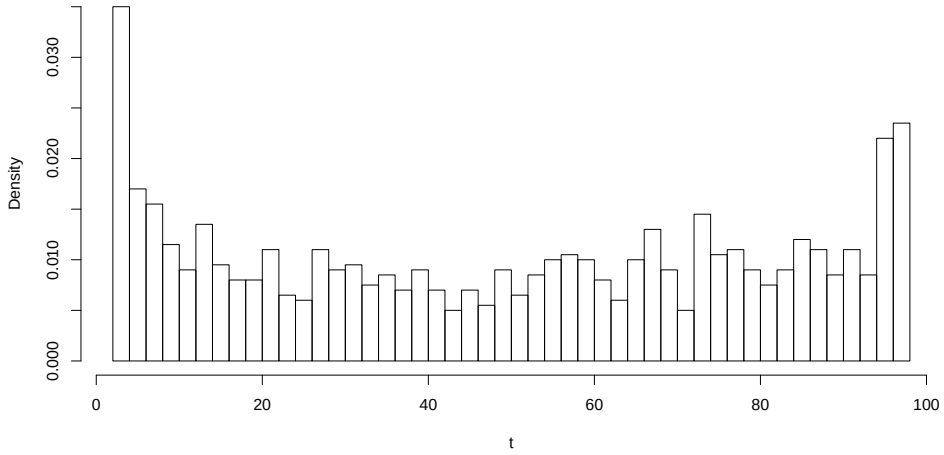


Figure 5.2: Histogram of the selected knot s based on the LR statistic for the one-knot CHBB model $M_{(1)}$. The data is generated from the null model $M_{(0)}$(a Brownian Bridge process with variance 1) and $T = 100$. The LR statistic are more likely to select a knot which is near the end points.

has $s-1$ observations. Under $M_{(0)}$ with $\varsigma^2 = 1$, $\hat{\sigma}^2_{1,1}$ follows a gamma distribution with shape $(s-1)/2$ and scale $2/(s-1)$, as shown in Appendix B.2. As seen in Figure B.1, the probability $Pr\{\hat{\sigma}^2_{1,1} < \varepsilon\}$ decreases with the increase of s, where $0 < \varepsilon < 1$ is an arbitrary small number. Now suppose we have two candidate models, $M_{(1)}$ and $M'_{(1)}$ with the same $\hat{\varsigma}^2, \hat{\sigma}^2_0, \hat{\sigma}^2_{1,2}$, but $s = 2$ in $M_{(1)}$ while $s = 20$ in $M'_{(1)}$. It is more likely that $M_{(1)}$ results in a smaller $\hat{\sigma}^2_{1,1}$ and thus a larger LR statistic D. Our model selection procedure will then select $M_{(1)}$, which has a very short period $s = 2$.

The sensitivity to small variance estimates is the main reason why model selection with the LR statistics tends to pick two knots that are very close to each other. This problem also carries over to the BFs, as the likelihood is a vital part of the BFs, especially the intrinsic and fractional BF. For small periods, the likelihood is not very informative about σ^2 and thus different priors can heavily affect the BF. The KL divergence has a similar problem. The KL for the above $M_{(0)}$ and $M_{(1)}$ is

derived in Appendix B:

$$\mathrm{KL}(M_{(0)}||M_{(1)}) = g\left(\frac{\hat{\varsigma}^2}{\hat{\sigma}_0^2}\right) + (s-1)g\left(\frac{\hat{\varsigma}^2}{\hat{\sigma}_{1,1}^2}\right) + (T-s-1)g\left(\frac{\hat{\varsigma}^2}{\hat{\sigma}_{1,2}^2}\right),$$

where $g(x) = x - \log(x) - 1$ with $g(x)$ approaching infinity when x approaches either 0 or ∞. The KL divergence is also very big when any one of $\hat{\sigma}_0^2, \hat{\sigma}_{1,1}^2, \hat{\sigma}_{1,2}^2$ is close to zero.

5.3 Reflections and future directions

The idea that led to the CHBB was inspired by the short period updates in our Bayesian Melding approach in Chapter 4. Our attempts may be viewed as trying to partition the GPS observations into different periods and to estimate a different variance parameter in each period. This might produce better predictions if we knew where to partition the GPS observations instead of selecting the partitions using the data. Yet, with such information, analyzing each period separately might work equally well. For the animal tracking data, it is possible to select the periods based on other information such as the diving records, or simply a visual examination of the data. Such selection methods are not practical for other spatial or spatio–temporal data sets, as there is usually no auxiliary information or it is extremely difficult to perform a visual examination. So it is necessary to have an automatic model selection procedure.

The problem with our current model selection procedure is that it tends to select short periods with small variance parameters. For the marine mammal tracking data, it results in too many behavior changing points. Also, we cannot tell the animal's different hidden behavior from the variance parameters, because they are all small. This disagrees with our initial expectation that a CHBB would have periods with small variances and other periods with large variances as in Figure 5.1. This might be fixed by designing informative priors on the σ^2s or priors on the knot structures similar to Bayesian trees (Chipman, 1998; Gramacy and Lee, 2008). Another direction would be to discard all the metrics above and design a metric that encourages $\sigma_{1,1}^2, \sigma_{1,2}^2, \ldots$ to be different. These ideas would require a lot of trial–and–error, which is part of our future work.

Another crucial difficulty in performing model selection for CHBB or CHGP is that the number of models (i.e., different knot sets) is 2^n for data observed at n locations. Such a number is prohibitive if we try to enumerate all models and find the optimal model even for moderate n. We avoid this issue in this initial investigation by performing a sequential greedy search and not revisiting the added/deleted knots. Yet, there is no guarantee that such a greedy search would converge to the optimal model.

We start the study of the CHGP by working on the CHBB, because the BB seems to be a good model in our application to animal tracking and it has a simple mathematical form that enables us to derive many statistics explicitly. However, there is only one parameter, the variance parameter, in the BB. As seen in Section 8.2, the variance parameter estimates can be affected by model misspecification and the sampling rate, which might not be a good indicator of changes in conditional dependence. For future developments of CHGP, we should start with a process that has correlation parameters and a predetermined knot structure.

Chapter 6

Generalized Ornstein–Ulhenbeck Process and Its Application in Animal Tracking

In Chapter 4, we showed that our Bayesian Melding (BM) approach with the Brownian Bridge process works well to combine the GPS observations and DR path for accurate high resolution track estimation. However, the constant variance σ_H^2 in the Brownian Bridge is not able to reflect the inhomogeneous feature of the track. Attempting to change only σ_H^2 does not lead to a successful model in Chapter 5. In this chapter, we propose a generalized Ornstein–Ulhenbeck process (GOU) to replace the Brownian Bridge process. GOU offers a flexible structure for both the mean and variance of the path. Unlike the "jumping" variances in CHBB, we control the mean and covariance of GOU to vary smoothly through our formulation based on B-splines, which also helps GOU retain a reasonably parsimonious parameterization. In addition, a GOU is still a Gaussian Markov process, which enables efficient computation via the Kalman filter and smoother.

Our definition of the GOU can be easily modified for a multivariate process, but we focus on a univariate process in this chapter. The rest of this chapter is organized as follows. Section 6.1 introduces the definition of the GOU process and its properties. Its application in our Bayesian Melding framework for combining the GPS observations and DR path for animal tracking is discussed in Section 6.2.

6.1 Generalized Ornstein–Ulhenbeck process

As reviewed in Section 3.2, only a few types of SDEs have closed form solutions and the Ornstein–Ulhenbeck (OU) process is one of them,

$$d\eta(t) = \rho(\mu - \eta(t))dt + \sigma dW(t),$$

with $\sigma > 0$. The initial value $\eta_0 = \eta(0)$ is considered to be fixed in our study. Using Itô's lemma (see e.g., Arnold, 1974), the solution of the above SDE is

$$\eta(t) = \eta_0 e^{-\rho t} + \mu(1 - e^{-\rho t}) + \sigma \int_0^t e^{-\rho(t-u)} dW(u).$$

In addition, $\{\eta(\cdot)\}$ can be shown to be a Gaussian process, whose mean and covariance are

$$\begin{aligned} \mathrm{E}\,\eta(t) =& \eta_0 e^{-\rho t} + \mu(1 - e^{-\rho t}) \\ \mathrm{Cov}\,(\eta(s), \eta(t)) =& \frac{\sigma^2}{2\rho} e^{-\rho(s+t)}(e^{2\rho \min(s,t)} - 1). \end{aligned}$$

If $\rho > 0$, $\eta(t)$ has a stationary distribution of $\mathrm{N}(\mu, \frac{\sigma^2}{2\rho})$ and has a mean–reverting behavior: when $\eta(t) > \mu$, $\eta(t)$ is expected to decrease to μ, when $\eta(t) < \mu$, $\eta(t)$ is expected to increase to μ. This behavior is illustrated in Figure 6.1 with two different initial values of $\eta_0 = 0, 10$ and $\mu = 5$. Notice that the $\rho > 0$ condition is only required to ensure the existence of a stationary distribution and the mean-reverting behavior, while the definition of the OU process does not require $\rho > 0$. When $\rho < 0$, $\sigma^2/(2\rho) < 0$ and $e^{2\rho \min(s,t)} < 1$ still results in a positive variance. In this case, the mean and variance will diverge to infinity as $t \to \infty$, when $\mu - \eta_0 \neq 0$. When $\rho = 0$, $\lim_{\rho \to 0}(e^{2\rho \min(s,t)} - 1)/(2\rho) = \min(s,t)$ and the OU process reduces to a Brownian Motion process with variance σ^2.

The OU process has desirable properties and therefore has been widely used in financial models for interest rates and stock prices (see e.g., Chan et al., 1992). Zhu et al. (2011a,b) applied the OU and integrated OU processes to model longitudinal data. But the mean function of the OU process is an exponential curve $\eta_0 e^{-\rho t} + \mu(1 - e^{-\rho t})$ as shown by the dashed curves in Figure 6.1, which is not sufficiently

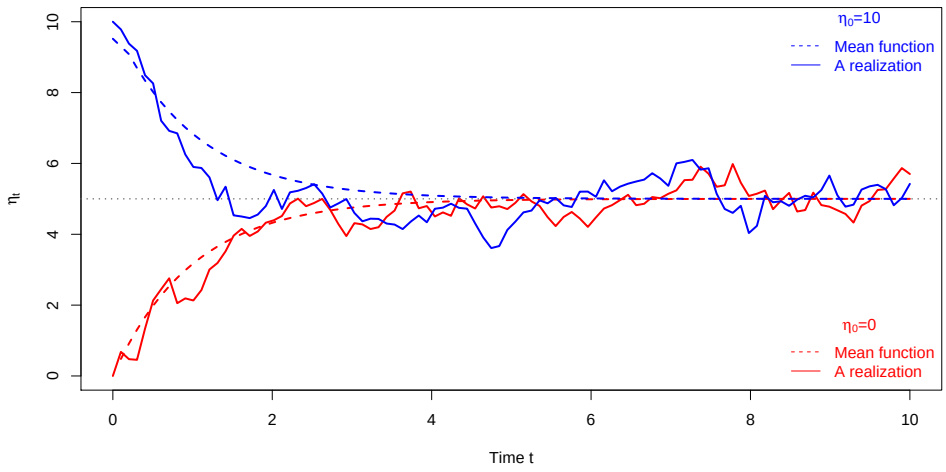


Figure 6.1: Examples of the Ornstein–Ulhenbeck process of $\rho = 1, \sigma = 1, \mu = 5$ with different initial value η_0. The mean curve is plotted with the dash lines while the realizations are plotted with the solid lines.

flexible in modeling the tracks. A natural extension of the OU process is the linear SDE as discussed in Chapter 8 of Arnold (1974),

$$d\eta(t) = (A(t)\eta(t) + a(t))dt + \sigma(t)dW(t). \tag{6.1}$$

Notice that the linear SDE reduces to a OU process when $A(t) = -\rho, a(t) = \rho\mu, \sigma(t) = \sigma$. When $A(t) = 1/(1-t)$, $a(t) = 0$, and $\sigma(t) = 1$, the linear SDE results in a Brownian Bridge process defined on $[0,1]$ (see e.g., Chapter IV of Rogers and Williams, 2000). In general, $A(t), a(t), \sigma(t)$ need to satisfy a few mild conditions listed in Section 3.2 to guarantee the existence and uniqueness of the solution. The mean and covariance functions of the linear SDE are derived in Section 6.1.1. We consider only the linear SDE instead of other types of SDE, as the linear SDE is often used as an approximation base for many other types of SDE (Ozaki, 1992; Vrettas et al., 2015). So even if we modeled with a non-linear SDE, we may be forced to use the linear one as an approximation in the inference and prediction unless we consider particle based Monte Carlo methods, such as the sequential Monte Carlo (Del Moral et al., 2006) or particle Markov Chain Monte Carlo (Andrieu et al., 2010), which are computationally intensive.

Although the linear SDE has been studied extensively in probability theory, we found few applications with this process except for the OU special case. In our study, we have merely investigated the special case that $\sigma(t)$ is a constant function while $A(t)$ and $a(t)$ are modeled with basis functions,

$$A(t) = \sum_{j=1}^{J_A} \gamma_j^{(A)} B'_{j,k_A}(t) \tag{6.2}$$

$$a(t) = \sum_{j=1}^{J_a} \gamma_j^{(a)} B_{j,k_a}(t). \tag{6.3}$$

Both $A(t)$ and $a(t)$ have equally spaced knots in the time domain. We choose the B-spline function for an initial investigation but other differentiable basis functions, such as other splines or Fourier basis functions may well be suitable. Notice that $A(t)$ is modeled by the first derivative of the basis functions. As reviewed in Section 3.4, the first derivative of the B-spline(J,k) can be represented by a B-spline$(J-1,k-1)$ with one less degree of freedom. The modeling of $A(t)$ via $B'_{j,k_A}(t), j = 1,2,\ldots,J_A-1$ is equivalent to using $B_{j,k_A-1}(t), j = 1,2,\ldots,J_A-1$. As explained later, using $B'_{j,k_A}(t)$ helps to simplify the evaluation of the linear SDE. We fix $\gamma_{J_A}^{(A)} = 0$ in this investigation to ensure the identifiability of $\gamma_j^{(A)}, j = 1,2,\ldots,J_A-1$.

The idea of modeling $A(t)$ and $a(t)$ through the B-splines allows the process to have flexible (non-stationary) mean and variance functions. We can easily control the level of flexibility by tuning the number of basis functions J_A, J_a. For notational simplicity, we call such a linear SDE with coefficients modeled by basis function as a "generalized Ornstein–Ulhenbeck process" (GOU). As discussed later, GOU remains a Gaussian Markov process whose computation can be handled efficiently with sparse matrix or the Kalman smoother techniques.

6.1.1 Mean and covariance of GOU: our model choice

As seen in Chapter 8 of Arnold (1974), the solution of (6.1) can written in the following form

$$\eta(t) = \Psi(t)\left(\eta_0 + \int_0^t \Psi(u)^{-1}a(u)du + \int_0^t \Psi(u)^{-1}\sigma(u)dW(u)\right),$$

where

$$\Psi(t) = \exp\left\{\int_0^t A(u)du\right\} = \exp\left\{\sum_{j=1}^{J_A}\gamma_h^{(A)}(B_{j,k_A}(t) - B_{j,k_A}(0))\right\}. \tag{6.4}$$

Notice that $\Psi(u)^{-1} = 1/\Psi(u)$, not the inverse function of $\Psi(u)$. When $A(t), a(t)$ satisfy conditions discussed in Section 3.2, which can be met by simply constraining the spline coefficients to be bounded in our formulation (6.2) and (6.3), the above solution is a Gaussian process with mean function

$$m(t) = \mathrm{E}\,\eta(t) = \Psi(t)\left(\eta_0 + \int_0^t \Psi(u)^{-1}a(u)du\right), \tag{6.5}$$

and covariance function

$$K(s,t) = \mathrm{Cov}\,(\eta(s),\eta(t)) = \Psi(s)\left[\int_0^{\min(s,t)} \sigma(u)^2\Psi(u)^{-2}du\right]\Psi(t). \tag{6.6}$$

In the above mean and covariance functions, we notice that the function A needs to be integrated first to calculate $\Psi(t)$, which in turn needs to be integrated again for the mean or covariance function. Modeling $A(\cdot)$ with a function whose integral can be calculated in close form, such as the B-spline or its first derivative, can help us avoid calculating both layers of integral numerically. We choose to use the first derivative, because its integral, namely the B-spline, can be easily calculated in software packages, such as the R package "fda" (Ramsay et al., 2014). This helps us to simplify the implementation of GOU. If we choose to model A with the B-spline, the integral expression needs to be manually derived.

The covariance function of GOU in (6.6) depends on both $\Psi(t)$ and $\sigma(t)$. To allow for a dynamically changing variance in the GOU, it is sufficient to let one

of $A(t)$ and $\sigma(t)$ vary. $\sigma(t) = \sigma$ is fixed as a constant hereafter to avoid overly complex parameterization of the GOU.

6.1.2 Conditional distribution of GOU

As seen in Chapter 8 of Arnold (1974), the conditional distribution $\eta(t)|\eta(s) = x$, $t > s$ is Gaussian with mean and variance respectively,

$$\mathrm{E}(\eta(t)|\eta(s) = x) = \Psi(t) \int_s^t \Psi(u)^{-1} a(u) du + x \exp\left(\int_s^t A(u) du\right) \tag{6.7}$$

$$\mathrm{Var}(\eta(t)|\eta(s) = x) = \sigma^2 \Psi(t) \int_s^t \Psi(u)^{-2} du. \tag{6.8}$$

Notice that the conditional mean of $\eta(t)$ given $\eta(s)$ is a linear function of $\eta(s)$, which enables us to express the GOU process at a finite set of time points $\mathscr{T} = \{t_0 = 0, t_1, t_2, \ldots, t_n\}$ as a state model in a dynamic linear model (DLM) (3.15). For expository simplicity, define $\theta_i = \eta(t_i) - \mathrm{E}\,\eta(t_i)$, $\theta_{i-1} = \eta(t_{i-1}) - \mathrm{E}\,\eta(t_{i-1})$ (θ_i has mean zero, i.e., we remove the mean of the GOU first). Substituting t_i, t_{i-1} into (6.7) and (6.8), the conditional distribution of θ_i given $\theta_{i-1} = x$ is

$$\begin{aligned}
\mathrm{E}(\theta_i|\theta_{i-1} = x) &= x \exp\left(\int_{t_{i-1}}^{t_i} A(u) du\right) \\
\mathrm{Var}(\theta_i|\theta_{i-1} = x) &= \sigma^2 \Psi(t) \int_{t_{i-1}}^{t_i} \Psi(u)^{-2} du.
\end{aligned}$$

This is equivalent to the following linear transformation from θ_{i-1} to θ_i,

$$\theta_i = G_i \theta_{i-1} + u_i, \qquad u_i \sim \mathrm{N}(0, C_i), \tag{6.9}$$

with

$$\begin{aligned}
G_i &= \exp\left\{\int_{t_{i-1}}^{t_i} A(u) du\right\} = \exp\left\{\sum_{j=1}^{J_A} \gamma_h^{(A)} \left(B_{j,k_A}(t_i) - B_{j,k_A}(t_{i-1})\right)\right\}, \\
C_i &= \sigma^2 \Psi(t) \int_{t_{i-1}}^{t_i} \Psi(u)^{-2} du.
\end{aligned}$$

Equation 6.9 will be used in the next section as part of a DLM representation of our Bayesian Melding approach.

6.1.3 Sparse precision matrix of GOU

According to Arnold (1974), the solution to the linear SDE that satisfies conditions in (3.12) and (3.13) is a Gaussian Markov process. Our GOU process is a special case of the linear SDE and thus a Gaussian Markov process. As seen in Rue and Held (2005), a realization of a Gaussian Markov process at any finite collection of points has a multivariate Gaussian distribution whose precision matrix is sparse. We can derive its covariance and precision matrix as follows. For a finite set of time points $\mathscr{T} = \{t_0 = 0, t_1, t_2, \ldots, t_n\}$, $t_i < t_j$ for all $i < j$, we denote the random variable at these times points by $\eta_i = \eta(t_i)$. From (6.6), the covariance of $\eta_i, \eta_j, i < j$ is

$$\text{Cov}(\eta_i, \eta_j) = K(t_i, t_j) = h_i h_j \sum_{k=1}^{i} d_k,$$

where

$$d_i = \int_{t_{i-1}}^{t_i} \Psi(u)^{-2} du, \quad h_i = \sigma \Psi(t_i).$$

Now define the following matrices,

$$\begin{aligned} \mathbf{D} &= \text{Diag}(\{d_1, d_2, \ldots, d_n\}) \\ \mathbf{H} &= \text{Diag}(\{h_1, h_2, \ldots, h_n\}), \end{aligned}$$

and $\mathbf{L}$ as a lower-triangular matrix of 1's as shown below. The inverse of $\mathbf{L}$ is a sparse matrix where only the diagonal and lower off-diagonal elements are nonzero.

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix} \qquad \mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

Then covariance and precision matrices for the random vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ are:

$$\begin{aligned}\mathbf{\Sigma} =& \mathbf{HLDL}^T\mathbf{H} \\ \mathbf{Q} =& \mathbf{\Sigma}^{-1} = \mathbf{H}^{-1}\mathbf{L}^{-T}\mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{H}^{-1}.\end{aligned} \tag{6.10}$$

The precision matrix is the product of three diagonal matrices and two off-diagonal matrices, which is a tri-diagonal matrix, such that

$$\begin{aligned}
&Q_{i,i} = h_i^{-2}(d_i^{-1} + d_{i+1}^{-1}), && i < n \\
&Q_{i,i+1} = Q_{i+1,i} = -h_i^{-1}h_{i+1}^{-1}d_{i+1}^{-1} && i < n \\
&Q_{n,n} = h_n^{-2}d_n^{-1} && \\
&Q_{i,j} = 0, && |i-j| > 1.
\end{aligned}$$

6.1.4 Implementation of GOU in practice

We discussed two properties of the GOU process that enable its efficient computation above. With the conditional distribution and the DLM representation (6.9), the Kalman filter and smoother discussed in 3.3 can be used to calculate the likelihood and posterior for GOU. We can also consider the sparse matrix techniques discussed in 3.1.1, as the GOU process has a tri-diagonal precision matrix. A natural question is which approach is more efficient.

In theory, both the sparse matrix and Kalman filter/smoother approaches, if implemented properly, can have linear complexity $O(n)$ in evaluating the likelihood or posterior. Notice that the sparse matrix approach usually has super linear complexity, i.e., is larger than $O(n)$, because of the Cholesky decomposition step (Davis, 2006). But for our GOU process, the Cholesky decomposition can be calculated explicitly in linear time from (6.10).

Although these two have the same theoretical complexity, they still have some differences in practice depending on the implementation and which programming language is used. We implemented the DLM approach for the GOU with the R package "dlm" (Petris et al., 2009) and the sparse matrix approach with the R package "Matrix" (Bates and Maechler, 2016). The "dlm" package offers the Kalman

filter/smoother implementation in both the R and Fortran languages. A simulation was used to compare the three implementations, DLM with R (DLM-R), DLM with Fortran (DLM-F), and sparse matrix (SparseM), in terms of the computation time in evaluating different likelihoods involving the GOU. We only consider the likelihood evaluation because it is the most frequently used calculation in our applications, i.e., likelihood evaluation is performed many times in the numerical optimization for the posterior mode and Hessian matrix.

The first likelihood in this simulation was that of a realization (free from observation noise) of the GOU at its true parameter. We also considered our Bayesian Melding (BM) model with the GOU, as introduced in the next section, and it can be viewed as the GOU process with measurement error and other bias components. We considered the number of time points $T \in \{100, 200, 500, 1000\}$ as close to the number of GPS time points in our real data. Multiple parameter settings were considered but the conclusions were invariant to these settings. We only report on one setting copied from a real data fit. Table 6.1 summarizes the total computational time for 100 replicates of these likelihood evaluations.

Table 6.1: Computation times (in seconds) for three different approaches to evaluate the likelihood of models involving the GOU. T is the number of time points.

	Likelihood of GOU			Likelihood of BM with GOU		
T	DLM-R	DLM-F	SparseM	DLM-R	DLM-F	SparseM
100	1.89	0.28	0.73	1.69	0.32	2.44
200	3.41	0.51	0.96	3.23	0.45	2.80
500	7.59	0.50	1.25	7.41	0.67	3.40
1000	15.31	0.96	2.00	14.66	1.13	5.40

From Table 6.1, we can see that the DLM with Fortran is the clear winner of the three approaches. Its advantage is moderate for small n and the likelihood of GOU, but quite remarkable for large n and the complex BM model. It can be more than 10 times faster than the DLM-R and approximately 5 times faster than the sparse matrix implementation when $T = 1000$. We must emphasize that Table 6.1 is a comparison of software packages that we can use to implement our models. The advantage of the DLM with Fortran might be mainly due to the fact that it is

using a faster language.

Another advantage of using the DLM implementation is that the same code can work for $T = 100$ and big $T > 500,000$. If we were to use the sparse matrix approach, we would need to separate such a long sequence into short pieces due to memory issues and to update them separately. This idea would require a large amount of derivation and programing, i.e., similar to what we did in Appendix A for the BM models with the BB. The DLM implementation requires less detailed programing and thus becomes the only implementation we consider hereafter.

6.2 Application to Bayesian melding for animal's track

We consider use of the GOU process to replace the Brownian Bridge process in our Bayesian Melding approach to combine the GPS observations $\{Y(t_k)\}$ and DR path $\{X(t)\}$, namely,

$$\boldsymbol{\eta}(0:T) \sim \text{GOU}(\boldsymbol{\gamma}^{(A)}, \boldsymbol{\gamma}^{(a)}, \sigma) \tag{6.11}$$

$$Y(t_k)|\eta(t_k) \overset{\text{iid}}{\sim} \text{N}(\eta(t_k), \sigma_G^2), k = 0, 1, \ldots, K \tag{6.12}$$

$$X(t) = \eta(t) + h(t) + \xi(t). \tag{6.13}$$

As in Section 4.2, we consider a set of equally spaced time points $\mathscr{T} = \{0, 1, 2, \ldots, T\}$, but our model works for irregularly spaced time points as well. $X(t)$ denotes the DR path at time t. $Y(t_k)$ denotes the kth GPS observation and $\{t_k, k = 0, 1, \ldots, K\}$, which is a subset of $\mathscr{T}$. We consider the random bias process $\xi(t)$ as a Brownian Motion process with unknown variance σ_D^2. The measurement error in the GPS observations has a fixed variance $\sigma_G^2 = 0.0625$. The systematic bias in DR $h(t)$ is considered to be an unknown constant β, which is the best model according to our model selection analyses in Section 4.5.2. In addition, $\boldsymbol{\gamma}^{(A)} = \{\gamma_1^{(A)}, \gamma_2^{(A)}, \ldots, \gamma_{J_A-1}^{(A)}\}$ and $\boldsymbol{\gamma}^{(a)} = \{\gamma_1^{(a)}, \gamma_2^{(a)}, \ldots, \gamma_{J_a}^{(a)}\}$ denote all the unknown spline coefficients for $A(t)$ and $a(t)$ ($\gamma_{J_A}^{(A)} = 0$ being fixed). We represent all the unknown hyper-parameters with $\boldsymbol{\phi} = \{\boldsymbol{\gamma}^{(A)}, \boldsymbol{\gamma}^{(a)}, \sigma, \sigma_D^2, \beta\}$. The priors for the hyper-parameters are as follows:

$$[\boldsymbol{\gamma}^{(A)}] \propto 1, \qquad [\boldsymbol{\gamma}^{(a)}] \propto 1, \qquad [\beta] \propto 1,$$
$$[\log(\sigma)] \propto 1, \qquad [\log(\sigma_D^2)] \propto 1.$$

6.2.1 DLM of the BM approach

As seen in (6.9), the zero-mean GOU can be written as a state model in DLM. Similarly, Brownian Motion over the discrete time set $\mathscr{T} = \{0,1,2,\ldots,T\}$ can be formulated as

$$\xi_t = \xi_{t-1} + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma_D^2),$$

where $\xi_t = \xi(t)$ for notational simplicity. Given the hyper-parameter vector $\boldsymbol{\phi}$, we first remove the mean from the observations, i.e.,

$$\begin{aligned} \text{GPS: } Y^*(t) =& Y(t) - m(t) \\ \text{DR Path: } X^*(t) =& X(t) - m(t) - h(t), \end{aligned}$$

where $m(t)$ is the mean of the GOU process as in (6.5). Then define $\mathbf{O}_i = \{X^*(t_i), Y^*(t_i)\}^T$ and $\boldsymbol{\zeta}_i = \{\eta(t_i) - m(t_i), \xi_i\}^T$. For the time points that the GPS observation is not available, we code them by "not available" (NA), which will be automatically marginalized in the Kalman filter/smoother.

The Bayesian Melding model with GOU (BM-GOU hereafter), is then written as

$$\begin{aligned} \mathbf{O}_i =& \mathbf{F}_i \boldsymbol{\zeta}_i + \mathbf{v}_t & \mathbf{v}_t \sim& N_m(\mathbf{0}, \mathbf{V}_t) \\ \boldsymbol{\zeta}_i =& \mathbf{G}_i \boldsymbol{\zeta}_{i-1} + \mathbf{u}_t & \mathbf{u}_t \sim& N_p(\mathbf{0}, \mathbf{U}_t), \end{aligned}$$

with

$$\mathbf{F}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad \mathbf{G}_i = \begin{bmatrix} G_i = \exp(\int_{t_{i-1}}^{t_i} A(u)du) & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{V}_t = \begin{bmatrix} \sigma_G^2 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{U}_t = \begin{bmatrix} C_i = \sigma^2 \Psi(t) \int_{t_{i-1}}^{t_i} \Psi(u)^{-2} du & 0 \\ 0 & \sigma_D^2 \end{bmatrix}.$$

With this state–space representation of our DLM, the Kalman smoother can used to evaluate the following posterior distributions,

$$[\boldsymbol{\phi}|\mathbf{X}_G, \mathbf{Y}] \qquad [\boldsymbol{\eta}|\boldsymbol{\phi}, \mathbf{X}, \mathbf{Y}].$$

where the notation is the same as in Section 4.2: $\mathbf{X}$ for the DR path, $\mathbf{X}_G$ for the DR path at the GPS time points, $\mathbf{Y}$ for the GPS observations.

As is Section 4.2, the marginal posterior of $[\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}]$ is approximated by

$$\begin{aligned}[\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}] &= \int [\boldsymbol{\eta}|\boldsymbol{\phi},\mathbf{X},\mathbf{Y}][\boldsymbol{\phi}|\mathbf{X},\mathbf{Y}]d\boldsymbol{\phi} \\ &\approx \int [\boldsymbol{\eta}|\boldsymbol{\phi},\mathbf{X},\mathbf{Y}][\boldsymbol{\phi}|\mathbf{X}_G,\mathbf{Y}]d\boldsymbol{\phi}, \end{aligned} \tag{6.14}$$

where the integral in the second line is calculated via numerical integration discussed in Appendix A.6. One may notice that the dimension of the hyper-parameters is much larger with the GOU than with the Brownian Bridge, which leads to a huge number of grid points if we perform the grid search in the direction of each eigenvector. For example, $J_A = 3, J_a = 3$ already leads to eight hyper-parameters (2 $\gamma_j^{(A)}$s, 3 $\gamma_h^{(a)}$s, σ, σ_D^2, and β) and 5 points in each dimension results in $5^8 = 390625$ total grid points. We avoid this issue by searching only in the direction of the first few eigenvectors that explain more than 95% of the variation in the hyper-parameters. This is acceptable because the purpose of the grids is to explore the likelihood surface.

6.2.2 Model selection for BM-GOU and comparison with other approaches

As in Section 4.2, we use leave–5–out cross validation (L5OCV) to perform model comparison and selection. For the B-spline models in $A(t)$ and $a(t)$ used in the GOU, we considered B-splines of order k_a, K_A $\{2,3,4\}$ (piecewise linear to cubic) and the number of basis functions J_a, J_A in $\{k_a(k_A)+1,\ldots,6\}$, i.e., at least one knot besides the end points. This results in 81 models in total. Due to the large number of models and the complexity in evaluating the GOU, we screened those models based on the CV-RMSE of the empirical Bayes posterior

$$[\boldsymbol{\eta}|\mathbf{X},\mathbf{Y}] \approx [\boldsymbol{\eta}|\hat{\boldsymbol{\phi}},\mathbf{X},\mathbf{Y}],$$

where $\hat{\boldsymbol{\phi}}$ is the posterior mode for $\boldsymbol{\phi}$. For the ten models with the smallest CV-RMSE in the empirical Bayes posterior, we further calculated their fully Bayesian posterior (6.14), whose CV-RMSE was used to decide the final model.

We report the model with the smallest CV-RMSE in the full Bayesian version in Table 6.2 with its model specification J_A, k_A, J_a, k_a and the cross validation performance. To compare the BM-GOU with the models we developed in Chapter 4, Table 6.2 also includes the CV-RMSE and coverage percentage from the selected Bayesian Melding model with the Brownian Bridge (BM-BB for short) and the downscaling model, which were shown before as the last two columns of Table 4.4.

Table 6.2: The best GOU model, RMSEs and actual coverage percentages of 95% credible intervals (in gray backgrounds) for the Bayesian Melding approach with the GOU process (BM-GOU). The CV-RMSE and coverage from the Bayesian Melding with the Brownian Bridge (BM-BB), and the Downscaling approach (DS) are also included from comparison.

	Best GOU				CV-RMSE and coverage percentages					
	J_A	k_A	J_a	k_a	BM-GOU		BM-BB		DS	
Trip 1 Northing	5	2	6	4	0.80	95.2	0.80	94.9	0.95	95.6
Trip 1 Easting	6	4	4	2	0.68	97.4	0.75	97.8	0.75	98.9
Trip 2 Northing	6	2	6	3	1.83	93.8	3.06	93.0	2.59	93.0
Trip 2 Easting	3	2	5	3	2.58	93.0	2.62	96.9	2.56	98.4

From the first four columns of Table 6.2, the best models for the four data sets have $J > 4$ in at least one of $A(t)$ and $a(t)$, so that we have a flexible component for the animal's track modeling. These shows that the GOU modeling can be useful in predicting the animal's track. When compared to the BM-BB and downscaling approaches, the BM-GOU performs at least as well. In particular, the BM-GOU is better than the BM-BB in three out of the four data sets and they are tied for the Trip 1 Northing. In Trip 2 Northing, the BM with GOU manages to reduce the CV-RMSE by 40% when compared to the BM-BB. The actual coverage percentage from the credible intervals for BM-GOU is also close to the nominal level.

To further compare the BM-GOU and BM-BB, we plot the posterior mean and credible intervals from these two approaches for the Trip 2 Northing data set in Figure 6.2. Similar plots are obtained in other data sets and therefore omitted. As in the BM-BB, the posterior mean for BM-GOU lies close to the GPS observations and indicates a good bias correction of the DR path.

The close-up of the middle part of this trip provided in Figure 6.3 shows that the

Table 6.3: The average width of posterior credible intervals from the BM-GOU and the BM-BB.

	Trip 1		Trip 2	
	Northing	Easting	Northing	Easting
BM-GOU	1.58	1.64	3.78	4.23
BM-BB	1.62	1.90	5.42	4.82

BM-GOU provides a smoother posterior mean and narrower posterior CI than the BM-BB. It is plausible that the BM-GOU can capture a more complex correlation structure in the process and thus enable greater smoothing. The BM-GOU borrows strength from neighboring DR observations to reduce the amount of uncertainty in the prediction. This results in the narrower CIs as in Table 6.3, which summarizes the average CIs width in the four data sets. Combining the results in Table 6.2 and 6.3, we find that the BM-GOU provides narrower CIs than the BM-BB without sacrificing the coverage percentage, another advantage of the BM-GOU over BM-BB.

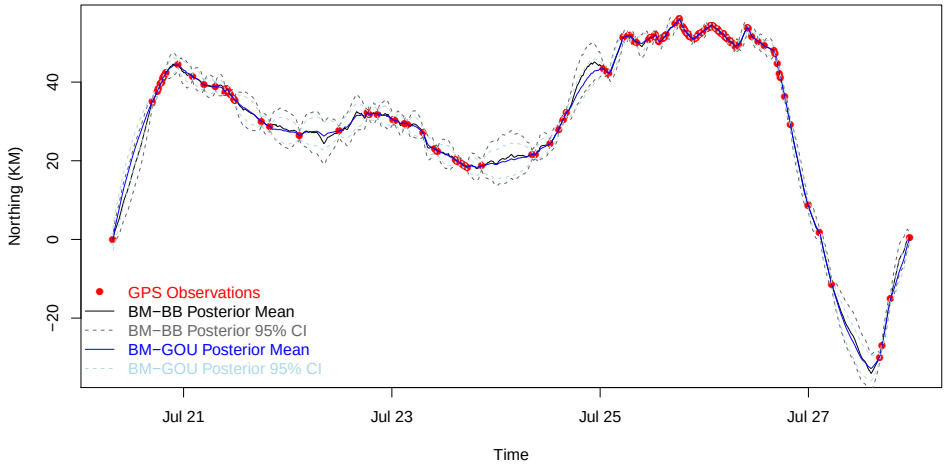


Figure 6.2: The posterior mean and 95% credible intervals from Bayesian Melding with GOU and Brownian Bridge for Trip 2 Northing data set.

In summary, based on our cross–validation study, we find that the flexible mean and covariance structure of the GOU process can improve the prediction power of

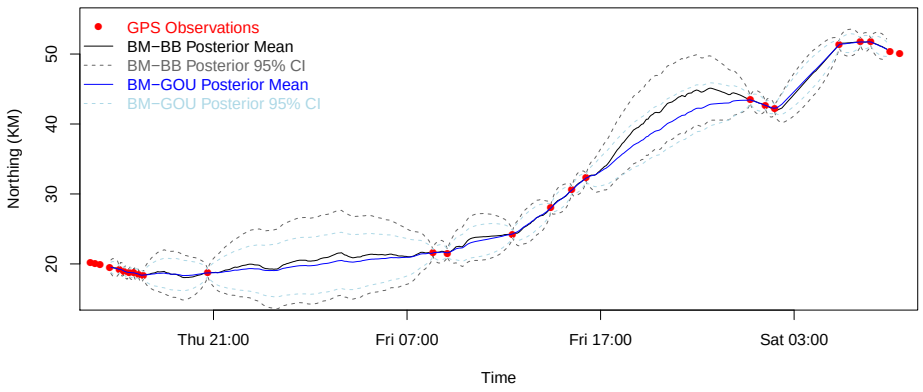


Figure 6.3: The posterior mean and 95% credible intervals from Bayesian Melding with GOU and Brownian Bridge for a period Trip 2 Northing data set.

Bayesian Melding approach. In addition, the model selection for GOU is much easier than that for the CHBB in Chapter 5, as the number of models for GOU is controlled by the number of basis functions in the splines while the CHBB creates 2^n models (n is the number of observations). This is a merit of modeling non-stationarity in a continuous fashion over the discrete fashion as in the CHBB. However, we studied the posterior for $\boldsymbol{\gamma}^{(A)}, \boldsymbol{\gamma}_h^{(a)}$ and did not find much useful information that can help to further interpret the animal's track. This motivates us to consider other forms of SDEs in the next chapter.

Chapter 7

Potential Field Modeling in Tracking

In the previous chapter, we demonstrated that SDEs can improve the prediction of our animals' spatial tracks when compared to the Brownian Bridge. In this chapter, we introduce more features into SDE modeling so that the fitted SDE can also be used to interpret the spatial tracks.

We work with a special form of the SDE with drift term formulated as the negative gradient of a function H:

$$d\mathbf{r}(t) = -\nabla H(\mathbf{r}(t),t;\boldsymbol{\beta})dt + \boldsymbol{\Sigma}^{1/2}(\mathbf{r}(t),t;\boldsymbol{\beta})d\mathbf{W}(t), \tag{7.1}$$

where $\mathbf{r}(t)$ denotes the location on the track at time t in $\mathbb{R}^d$. The error term $\mathbf{W}(t)$ is a d-dimensional standard Brownian motion with independent increments in each dimension and $\boldsymbol{\Sigma}(\mathbf{r}(t),t)$ is a covariance matrix. We fix it to be a diagonal matrix $\boldsymbol{\Sigma}(\mathbf{r}(t),t) = \sigma^2\mathbf{I}_d$ for the following two case studies.

The key feature of (7.1) lies in the drift term $-\nabla H(\mathbf{r}(t),t)$. Function $H(\cdot,\cdot;\boldsymbol{\beta})$ can vary with location and time and depends on other parameters $\boldsymbol{\beta}$. Its gradient $\nabla H = (\partial H/\partial x, \partial H/\partial y, \ldots)^T$ decides the negative expectation of the velocity of the object, such that the object is expected to visit locations of low values of H. This formulation comes from physics originally. Imagine that H represents the height of a hill and a ball is dropped on top of this hill. Obviously, the ball will

move down the hill, which is in the direction of locations with lower values of H. Brillinger et al. (2004), Brillinger et al. (2007), and Brillinger and Stewart (2010) introduced this approach to modeling the movement of various objects, such as elk, monk seal, and soccer ball. Russell et al. (2016) extended the potential field model by including a semi-parametric potential surface and a separate motility surface. Hooten et al. (2017) reviewed similar models developed for animal telemetry data. The function H is called the potential field (PF) in these papers, as $H(\mathbf{r},t)$ presents the potential that the object will move to location $\mathbf{r}$ at time t. A lower value of H indicates a higher likelihood of being visited.

The PF function H offers a versatile interface for modeling the tracks. In addition to time and location, it can also include covariates, such as the locations of other animals (Brillinger et al., 2007) or a map of an ocean's temperature or salinity. The coefficients of the covariates help us understand how an object interacts with others or the environment. Even without covariates, the PF can still offer an intuitive map to highlight attractive areas in the field of movement, which will be shown later in this chapter.

As discussed in Section 3.2, the solution to (7.1) may not have a closed form. We use the Euler approximation for inference:

$$\boldsymbol{r}(t_{i+1}) - \boldsymbol{r}(t_i) \approx -\nabla H(\boldsymbol{r}(t_i), t_k; \boldsymbol{\beta})(t_{k+1} - t_k) + \boldsymbol{\Sigma}^{1/2}(\boldsymbol{r}(t_k), t_k; \boldsymbol{\beta})\sqrt{t_{k+1} - t_k}\boldsymbol{\varepsilon}_k,$$

where $\boldsymbol{\varepsilon}_k, k = 0, 1, 2, \ldots, K$ are d-dimensional white noise terms (independent in each dimension). After this approximation, the parameter $\boldsymbol{\beta}$ can be estimated via maximum likelihood. When the PF $H(\cdot,\cdot;\boldsymbol{\beta})$ is a linear function of $\boldsymbol{\beta}$ and the diffusion matrix $\boldsymbol{\Sigma}^{1/2}(\cdot,\cdot;\boldsymbol{\beta})$ is a constant that does not depend on either the location $\mathbf{r}$ or the coefficients $\boldsymbol{\beta}$, the maximum likelihood estimate can be calculated by least squares directly.

In this chapter, we further develop the PF approach for two applications, the tracks of basketball movement in the National Basketball Association (NBA) games as well as the foraging paths of northern fur seals. They are discussed in Section 7.1 and 7.2 respectively.

7.1 Potential field for learning basketball strategies

Since the beginning of the 2013-2014 season, the NBA has used SportVU®[1] technology to track the movement of the players and the basketball in a nearly continuous manner. In each NBA arena, six high speed cameras have been installed to record images of each game at a rate of 25 readings per second. These images are then processed by specialized software to obtain the players' and ball's locations at the same rate. This technology has already yielded an array of informative statistics such as speed, distance, and touches. All these extend the ability to summarize a player's or team's performance, thereby allowing teams to assess and optimize strategy. For example, this technology was used recently to help Duke University win the 2014-2015 NCAA basketball championship [2].

The player tracking data encodes a vast amount of information for each game and has sparked a lot of innovative statistical research. For example, Cervone et al. (2014) modeled basketball possession using a Markov chain and derived the expected scores from this possession[3] to evaluate the efficiency of players' decisions. Cervone et al. (2016) further studied the impact of space occupation via hierarchical spatial models on both the offensive and defensive ends (i.e., expected points scored or defended resulted from the space occupation). Hidden Markov models and log Gaussian Cox processes were recently used to quantify players' defensive skills and effectiveness in Franks et al. (2015). McIntyre et al. (2016) developed a multinomial logistic regression model to identify the ball screens[4] and then applied the identified screens to performance assessment. Neural network models were used to classify different types of offensive play in Wang and Zemel (2016).

Many useful new statistics have arisen from the above cited studies to quantify players' or teams' performance, but few of them have aimed at learning a team/-player's playing strategies directly from the massive tracking data. The supervised learning approaches developed in McIntyre et al. (2016) as well as Wang and Zemel (2016) can learn certain types of plays, but they require human-annotated data sets

[1] http://www.stats.com/sportvu/sportvu-basketball-media

[2] http://www.stats.com/press-releases/duke-utilizes-stats-sportvu-player-tracking-technology-championship-season/

[3] A possession is the period when a player or a team is in control of the basketball.

[4] A screen is a basketball tactic that one player uses his/her body to block the defender(s) and thus create open space for his/her teammate.

to train the classifier (logistic or neutral network), which can be hard to obtain and prone to labeling errors. It is also difficult for the models learned in this fashion to adapt to the new types of plays created in the fast–evolving nature of NBA games. That leads us to consider an unsupervised learning approach for the massive NBA tracking data set, such that the play strategies can be learned and visualized without using human-annotated training data sets. Our PF approach is designed as such an unsupervised learning approach.

In this section, we illustrate our approach with the NBA tracking data from the seven game series between the Houston Rockets (HOU) and Los Angeles Clippers (LAC) in the 2015 NBA Western conference semi-finals. The raw player tracking data were downloaded from *http://stats.nba.com* [5] via Python and further processed by R (R Development Core Team, 2008). The raw tracking data contains the team, player names and jersey number, game clock[6], shot clock[7], and the movements of the ball and the players on the court. We further supplemented these data with the "play-by-play" record from *http://nba.com* in order to extract the time periods when a certain team (HOU or LAC) was in possession of the ball. Two different PF models are introduced below for different analytical purposes.

7.1.1 Potential field as a tensor product spline over the space

First, we consider the PF as a function over the basketball court, such that some areas of the court can be more attractive to the ball than other areas. This spatial function is formulated as a tensor product spline (De Boor et al., 1978)

$$H(\boldsymbol{r}(t),t;\boldsymbol{\beta}) = \sum_{j=1}^{K}\sum_{k=1}^{K} \beta_{jk} S_j\big(x(t)\big) T_k\big(y(t)\big),$$

where S_j, T_k are spline functions in the x (long edge of the basketball court) and y (short edge of of the basketball court) directions. The location vector is $\mathbf{r}(t) = (x(t), y(t))^T$. We fit such a PF with $K = 5$ to the basketball's movement paths

[5] Data used in our study was downloaded in Sep 2015. Unfortunately, those data are no longer available to the public.

[6] It shows the time remained in the quarter of a basketball game.

[7] NBA rules specifies that the offensive team must shoot within 24 seconds from the start of this play. This clock indicates how many seconds are left to shoot the basketball.

from the offensive plays by HOU in Game 4 (G4) and Game 5 (G5) of this series. In G4, we have tracks of the basketball movement in 89 HOU offensive plays which consist of 18,255 records of (x,y) locations. In G5, there are 81 plays, corresponding to 15,346 moments. The fitted PF is plotted in Figure 7.1.

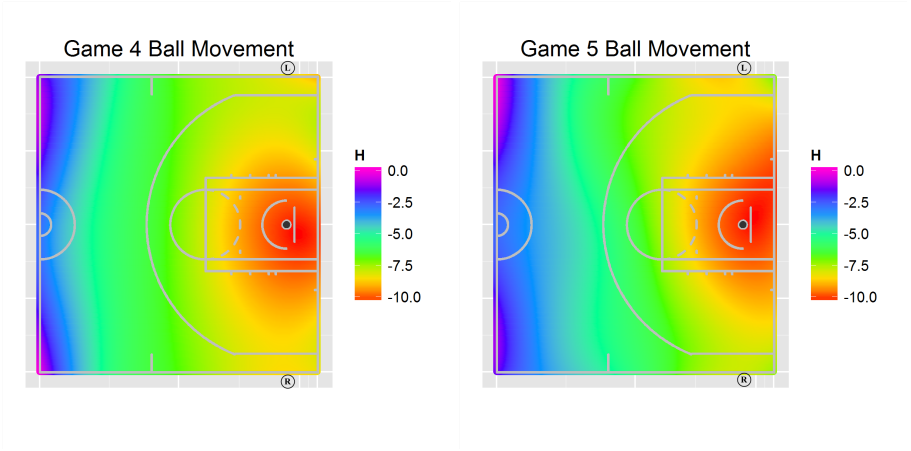


Figure 7.1: Potential fields parameterized by tensor product splines fitted to the paths of the ball when HOU was on offence during Games 4 and 5 vs. LAC in the 2015 NBA Conference playoffs. LAC won Game 4 128-95, while HOU won Game 5 124-103. The colors show the values of the potential field H, whose lowest value is around the basket.

At first, it appears that the two plots in Figure 7.1 are similar. In both games, we observe that the PFs take on high values around the half court line (blue) and display a tendency to steadily decrease toward the basket (green, yellow, then red). The minimum value of the PF is attained near the basket, as can be seen from the red area surrounding this location. This is plausible, as the ball is supposed to move toward the basket. Moreover, the PFs quickly decrease near the half court line because the player dribbling the ball up the court is likely to rush over the half court line to setup the offense (The gradient of the PF determines the velocity of the movement). The PFs for both games also indicate that Houston's usual offensive patterns are to move the ball inside or to move the ball to the corners for a 3-point shot[8]. This can easily be seen from the plot by observing the red and yellow

[8]Ball shot from somewhere outside the big arch (3-point line).

regions.

The same PF model can also be applied to analyze player's movement data. We work with the tracks of Jason Terry (point guard) and Trevor Ariza (small forward) from the HOU team in G5 as an illustration. In this game, we have 52 and 50 plays involving Trevor Ariza and Jason Terry, respectively. These are limited to possessions in which HOU was on offence while Trevor Ariza/Jason Terry was on the court. The paths of the two players, from the raw data, are shown in Figure 7.2. Different colors in this figure correspond to different plays. The paths of each play are plotted as dots, which increase in size as the play develops. That is, the size of the dot indicates how much time has elapsed on the shot clock, with larger dots indicating less time remaining.

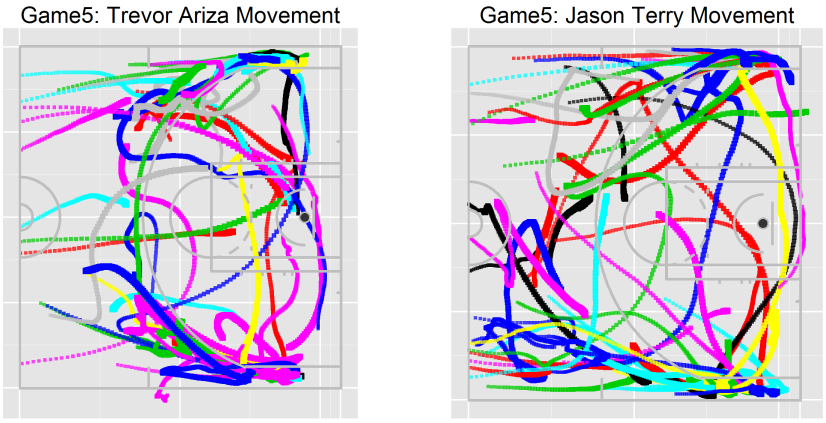


Figure 7.2: Movement paths for Trevor Ariza (left) and Jason Terry (right) in Game 5. The points increase in size as the play progresses (and the shot clock decreases). Different colors corresponds to different plays.

From Figure 7.2, we see that the paths of these two players typically end up along the sidelines, just beyond the three point lines in the corners. However, given the amount of overlap, it is difficult to infer anything more from Figure 7.2. On the other hand, the fitted PFs for these two players (Figure 7.3) display some interesting trends. The PF for Ariza takes on low values in both corners, but with preference toward the right corner (R). Conversely, Terry seems to prefer the left corner (L), indicated by the red region at the top left corner of the plot. It is a good

strategy for them to occupy different corner, as this can create more opportunities for inside–out pass.

We also note that the PF for Ariza takes on relatively low values near the basket, at a similar level to the left corner. This indicates that moving toward the basket (either by cutting or driving to the basket) versus staying in the corner, are equally attractive options to Ariza. This pattern is not observed in Terry's PF. This agrees with the fact that Ariza can drive to the basketball more often than Terry, because Ariza is younger and more athletic.

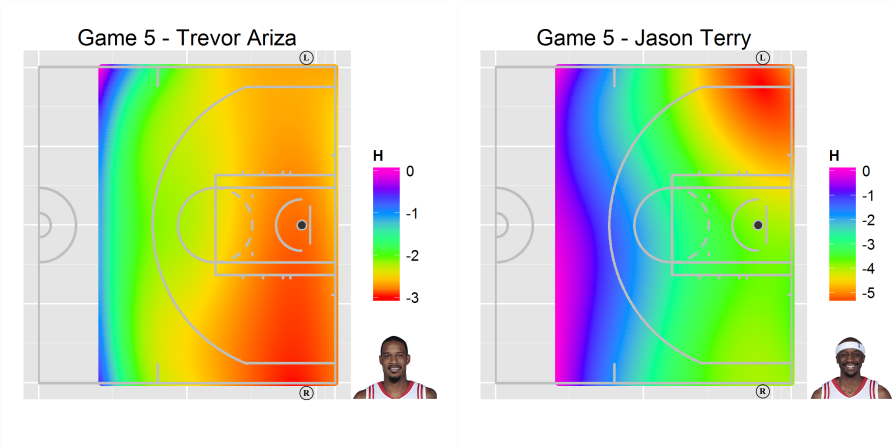


Figure 7.3: The fitted potential fields for Trevor Ariza (left) and Jason Terry (right) in Game 5. We choose to only show the potential fields away from the center circle, as to allow for clear visualization of the patterns in the potential fields. These two players tend to occupy different corners of the court.

7.1.2 Potential field using the other player's location as a covariate

The above PF with tensor product splines are useful to illustrate the patterns in ball or a player's movement. We can also use the PF approach to characterize how one player interacts with his four teammates by considering the following PF

$$H(\boldsymbol{r}(t),t;\boldsymbol{\beta}) = \beta_0 \|\boldsymbol{r}(t)_{\text{ball}} - \boldsymbol{r}_{\text{basket}}\|_2^2 + \sum_{j=1}^{4} \beta_j \|\boldsymbol{r}(t)_{\text{ball}} - \boldsymbol{r}(t)_{\text{player}_j}\|_2^2, \qquad (7.2)$$

where $||\cdot||_2$ denotes the l_2 norm (Euclidean distance). This PF model can be fitted to the moments when a certain player is in possession of the basketball. The different possessions of the ball by the same player even in the same play are assumed to be mutually independent. Namely, if the player passes the ball to one of his teammates and later gets the ball back, the two periods (possessions) are independent observations of the PF. The coefficient, β_0 represents the likelihood of the ball to move toward the basket, i.e., if $\beta_0 > 0$, the player is likely to move the ball toward the basket, especially if the distance between the ball and the basket is large. On the other hand, if $\beta_0 < 0$, then the player is less likely to move the ball toward the basket. The interpretations for β_j, for $j = 1, ..., 4$, are similar to β_0. These coefficients denote the likelihood of the player to move the ball toward teammate j.

We fit this model to HOU's starting lineup in G5, which consists of small forward Trevor Ariza, center Dwight Howard, shooting guard James Harden, point guard Jason Terry, and, power forward Josh Smith. We found 22 plays (3900 moments) by this line-up. We fit five separate models of (7.2) to the moments when each player is in possession of the ball. For expository simplicity, we denote the fitted coefficient for teammate j, or basket, when player i is in possession, by $\beta_{i,j}$ and therefore $\beta_{i,j}$ and $\beta_{i',j}$ are from different models when $i \neq i'$. For example, $\beta_{\text{Terry,Harden}}$ is the coefficient corresponding to Terry moving the ball toward Harden (via passing or dribbling). Similarly, $\beta_{\text{Harden,Basket}}$ is the coefficient corresponding to Harden moving the ball toward the basket. These coefficients are summarized in Table 7.1.

We begin by focusing on Houston's guard, Jason Terry. The negative coefficients for Terry to all of the players except for Harden and the basket indicate that Terry is unlikely to move the ball toward any teammates other than Harden. This corresponds to how HOU usually starts an offensive play: Terry brings the ball over the half court and passes to Harden to start the offense. James Harden has positive coefficients corresponding to Ariza and Terry as well as the basket. This agrees with what we observe by watching this game. Frequently, Harden would drive the ball to the basket, as indicated by the positive coefficient of 4.90; If he sees either Ariza or Terry open for a shot, he will make a pass to one of them. This tactic is effective because Harden has great slashing ability, often drawing defenders toward

Table 7.1: Fitted coefficients (multiplied by 1000) of the potential fields corresponding to HOU's most commonly used lineup in Game 5. They reflect the likelihood of the ball to move toward "column name" when "row name" is in possession of the ball.

	Ariza	Smith	Howard	Harden	Terry	Basket
Ariza	NA	-3.35	-1.51	1.90	1.03	5.38
Smith	1.14	NA	-1.23	-1.09	0.44	1.59
Howard	1.19	3.98	NA	-5.28	2.97	-6.27
Harden	0.72	-2.65	-3.70	NA	1.73	4.90
Terry	-1.21	-1.36	-3.04	2.29	NA	4.21

him. This often leaves one teammate open for a shot.

From the third row of Table 7.1, the likelihood of Howard moving the ball toward the basket is negative ($\beta_{\text{Howard,Basket}} = -6.27$). This seems counterintuitive, but is reasonable. In this game, Howard usually receives the ball near the basket and typically has three options: dunk, execute a post move[9], or pass back to the perimeter. The latter two options were more frequent, which results in the ball moving away from the basket and thus a negative coefficient.

Meanwhile, we admit that this formulation of PF for the interactions among player is not ideal. The likelihood of passing or driving is confounded with the distance between the players (or the basket), which leads to tricky cases in the interpretation. For example, imagine that Terry has already dribbled the ball close to the basket free of the defenders while Harden is still far away near the midcourt. Terry in this case will almost certainly shoot the basketball instead of pass to Harden. Yet, according to the PF model, Terry should be more likely to pass to Harden as he is far away (the distance from Terry to Harden is much larger than Terry to the basket). This is a shortcoming of our current model and should be addressed in future work. We may borrow ideas from the successful modeling of animal's social network as in Scharf et al. (2017) and the references within.

[9] Some moves/actions that a basketball player plays near the basket to get a better shot.

7.2 A mixture of potential fields model for seal foraging tracks

The PF model can also be used to interpret the foraging tracks of northern fur seals. As in the previous section, the fitted PF can serve as a map of hot foraging spots of the seals. For initial investigation, the PF is only fitted to the GPS observations in longitude and latitude without using the high resolution DR paths. As there are only approximately a few hundred GPS observations, we consider a simpler parameterization of the PF instead of the tensor product splines in the previous section:

$$H(\mathbf{r},t;\boldsymbol{\beta}) = \beta_1 x(t) + \beta_2 y(t) + \beta_3 x(t)^2 + \beta_4 y(t)^2 + \beta_5 x(t) y(t). \tag{7.3}$$

where $\mathbf{r}(t) = (x(t), y(t))^T$ are the longitude (Easting) and latitude (Northing) location.

This is also the PF model proposed for the movement of monk seals in Brillinger et al. (2007). Using Trip 2 as an example, the fitted PF (7.3) is shown in the top-left panel in Figure 7.4. The fitted PF does not offer a good interpretation of the animal's movement, as the area with the lowest PF is an area the animal circled around. This contradicts the interpretation of the PF, the area with lowest PF value should be most attractive to the animal. The poor fitting of this PF is plausible as it is impractical to assume that a single PF governs the whole path of the seal. For most of the trip, the seal aims to collect as much food as possible, so a PF of food resources would interpret the animal's movement. Once the animal has accumulated enough food, it needs to return to the island to feed its pup and the island becomes the only destination for the return part of this trip, which requires a different PF.

Brillinger et al. (2007) avoided this issue by manually chopping the trip into different periods. We attempt to tackle this issue by extending (7.3) into a mixture of PFs as follows:

$$H(x(t),y(t);\boldsymbol{\beta}) = \sum_{k=1}^{K} Z_k(t)\big(\beta_1^{(k)}x(t) + \beta_2^{(k)}y(t) + \beta_3^{(k)}x(t)^2 + \beta_4^{(k)}y(t)^2 + \beta_5^{(k)}x(t)y(t)\big),$$

where $\{Z_1(t), Z_2(t), \ldots, Z_K(t)\} \overset{\text{iid}}{\sim}$ from a multinomial distribution with $n = 1$ and a probability vector $\mathbf{p} = (p_1, p_2, \ldots, p_K)^T$. For simplicity, we assume the error terms in the longitude and latitude components are independent and have the same variance, such that $\boldsymbol{\Sigma}(\mathbf{r}(t),t) = \sigma^2\mathbf{I}_2$. Parameter estimates in the Euler approximated model can be calculated by the Expectation–Maximization (EM) algorithm of Dempster et al. (1977). For the current analysis, we only consider a two component mixture and calculate the parameter estimates with the FlexMix (Grün et al., 2008) package.

The mixture of PF models appears to be better than the original PF model, as the maximum of the mixture PF is in an area where the animal appeared to stay for a long period of time. The advantage of the mixture of PF model is more obvious if we study its two components separately, as in the bottom two panels of Figure 7.4. The maxima of the first PF are in two areas where the animal might be foraging as it went back and forth. The second PF has only one maximum at the island (start and end points of this trip). The first PF may reflect the food resources in the ocean, which attracts the foraging animals. The second PF reflects the animal's home, where it returns after foraging.

The above interpretation can be verified with the cluster membership[10] plotted in the top right panel. Those GPS observations close to their previous and next observations mostly belong to the first "foraging" component while those GPS observations being far from their previous and next observations are mostly in the second "returning home" component.

[10] Decided by the posterior probabilities from the EM algorithm

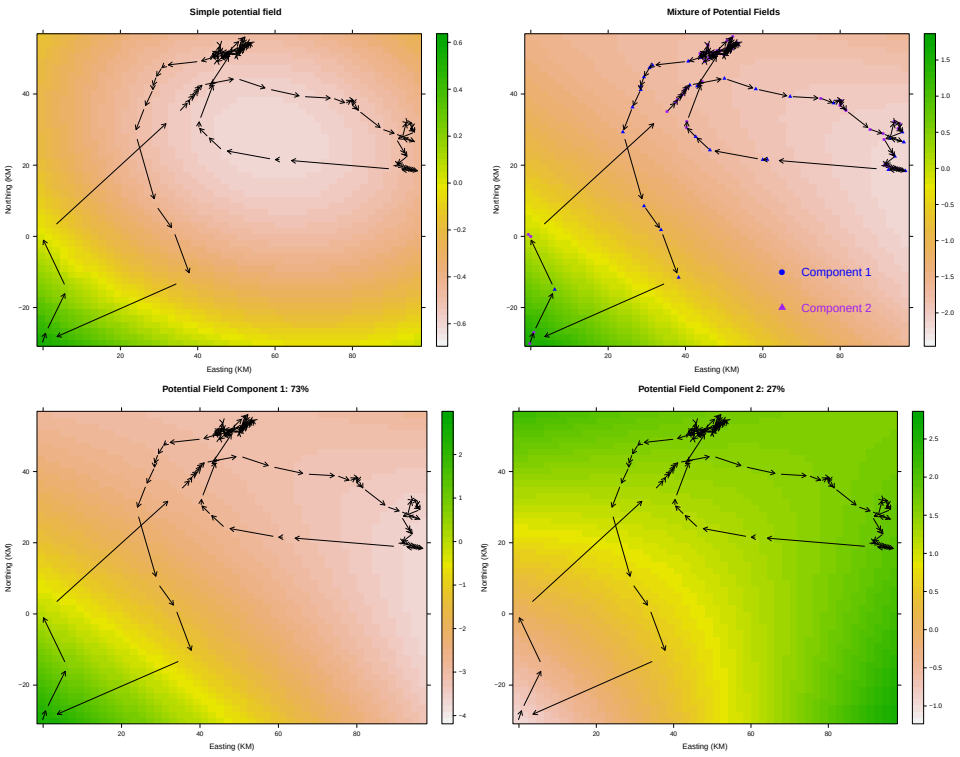


Figure 7.4: The simple potential field and mixture of potential fields models fitted for the Trip 2 GPS data set collected in our case study. Top left: the potential field with the original Brillinger model, which does not fit the track well; Top right: The mixture potential field and the posterior clusters; Bottom left: the first component of the mixture. It accounts for 73% of the observations and may be interpreted as the food map for the animal; Bottom right: second component of the mixture. It accounts for 27% of the observations and reflects the returning behavior of the animal. The top right field is the weighted sum of the bottom two fields with weight $0.73, 0.27$.

Chapter 8

Summary, Discussion, and Future Work

In the final chapter of this thesis, we first summarize our major contributions in the previous chapters in Section 8.1. Then we discuss a potential problem of GP modeling under different sampling frequency in Section 8.2. Some additional future work is discussed in Section 8.3.

8.1 Summary

In this thesis, we developed statistical models for high resolution tracking data, in particular, marine mammal tracking data with multiple sources of observations. Chapter 2 reviewed the scientific background of marine mammal tracking and studied the properties of the two distinctive observations: the accurate but sparse GPS observations and the high-resolution but biased Dead–Reckoning (DR) path. Such data from a case study of tracking of northern fur seals were used throughout this dissertation. The models we developed were built upon recent developments in spatio–temporal modeling with Gaussian processes (GP) and combined with various forms of stochastic differential equations (SDE), which were reviewed in Chapter 3 with special emphasis on how to make inference and predictions with them.

Chapter 4 first reviewed two Bayesian frameworks, Bayesian Melding (BM) and downscaling, commonly used to combine data from different sources for sta-

tistical inference. We adapted them to estimate the path of a moving object and applied them in our case study of tracking northern fur seals. To make the BM approach computationally feasible for big tracking data, we exploited the properties of the processes along with approximations to the likelihood to break the high dimensional problem into a series of lower dimensional problems. To implement the alternative, downscaling approach, we used the integrated nested Laplace approximation (INLA) to fit a linear mixed effect model that connects the two sources of observations. The predictions of the two approaches were compared by cross–validation as well as simulations. We showed that both approaches yield similar results—both provide accurate, high resolution estimates of tracks, as well as Bayesian credible intervals to characterize the uncertainty about the estimated paths. They outperformed the conventional bias correction of the DR path or interpolation methods based on GPS observation only in terms of prediction accuracy and coverage percentage.

Methods developed in Chapter 4 cannot model the inhomogeneous feature of the tracks. We first proposed the conditionally heterogeneous Gaussian process (CHGP) and the conditionally heterogeneous Brownian Bridge (CHBB) process in Chapter 5. Unfortunately, the idea of conditional heterogeneity produces too many candidate models and makes the model comparison metrics sensitive to small variance parameters. We failed to develop an adequate model selection method for the CHBB.

The generalized Ornstein–Ulhenbeck (GOU) process developed in Chapter 6 tackled the inhomogeneous issue from another angle. Through modeling the coefficients in a linear SDE via splines, we achieved a flexible non-stationary mean and covariance structure with parsimonious parametrization. With the Markovian property of the SDE, we were able to express the GOU as a dynamic linear model and thus perform its inference and prediction efficiently via the Kalman filter/smoother, which is also computationally scalable for big data. Our cross–validation studies show that the BM with GOU has better prediction accuracy than the BM and downscaling approaches from Chapter 4. The BM with GOU also achieved narrower credible intervals without sacrificing the coverage percentage.

To bring in more interpretable structures in the SDE modeling, we studied the potential field (PF) approach in Chapter 7. The PF is a SDE with its drift term

parameterized as the gradient of another function, which can be a function over the space or include other covariates. The PF was effective in learning and visualizing the trends in the tracks, which was illustrated by the marine mammal tracking data as well as the NBA tracking data. We also designed a mixture of PF's for the animal track to accommodate the different behaviors of the animal.

In summary, we developed many statistical models that can be used to predict and interpret tracking data. They were all designed to be scalable and efficient in computation. For the animal tracking data, the modeling approaches we developed and tested can facilitate the processing of high resolution in–situ records of the hydrographic data collected by marine mammals, and can contribute to broadening knowledge about parts of the ocean that have been hard to observe. They can contribute to studies seeking to address the effects of climate change on the ocean, and also contribute to answering many biological and ecological questions about habitat preference and resource selection (Hooten et al., 2013). The Bayesian data fusion approaches studied in this thesis can be further applied in other scenarios for combining data from different sources. The GOU and PF approaches that we developed can also be used to model other types of tracks.

In addition, we empirically assessed the models and identified best model choices based on cross–validation or visual examination in the applications considered in this study. The analyses demonstrated how users in other contexts may consider choosing models that best suit their applications, particularly where high dimensional data are involved.

8.2 Another examination of the approximation and model specification in Bayesian melding

In the Bayesian Melding approach developed in Chapter 4, one of the key approximations we utilized to simplify the posterior evaluation was

$$[\boldsymbol{\phi}|\mathbf{X},\mathbf{Y}] \approx [\boldsymbol{\phi}|\mathbf{X}_G,\mathbf{Y}]. \tag{8.1}$$

A similar approximation is also used in the Bayesian Melding with the GOU in Chapter 6. This approximation in (8.1) enables us to infer the two variance param-

eters $\boldsymbol{\phi} = (\sigma_H^2, \sigma_D^2)^T$ from those observations at the GPS time points instead of the whole data set. In our case studies, $\mathbf{X}$ on the left hand side of (8.1) is a vector of size over half a million, while $\mathbf{X}_G$ on the right hand side only has a couple of hundreds of elements. This is an approximation first designed for the computational issue and we have shown in Section 4.4.1 that it has little impact on the inference of the variance parameter based on *simulated* data. However, what happens when we remove this approximation and work with $[\boldsymbol{\phi}|\mathbf{X}, \mathbf{Y}]$ directly on our real data?

We do not have enough computing power to evaluate $[\boldsymbol{\phi}|\mathbf{X}, \mathbf{Y}]$ but we can mimic what will happen by thinning $\mathbf{X}$ to various resolutions. We thin the original 1Hz DR path at various rates $R = \infty, 600, 300, 120, 60$ and then combine them with the DR path at GPS time points. We denote such a data set by $\mathbf{X}_R \cup \mathbf{X}_G$. Thinning rate ∞ means that we only work with $[\boldsymbol{\phi}|\mathbf{X}_G, \mathbf{Y}]$. The smaller the rate the more observations from $\mathbf{X} \backslash \mathbf{X}_G$ are used in $[\boldsymbol{\phi}|\mathbf{X}_R \cup \mathbf{X}_G, \mathbf{Y}]$. Sparse matrices have to be used to make the computation feasible. The first two columns of Table 8.1 summarize the posterior modes of the two variance parameters.

To our surprise, the posterior modes of the variance parameters dramatically decrease when more of $\mathbf{X}$ is used in evaluating $[\boldsymbol{\phi}|\mathbf{X}_R \cup \mathbf{X}_G, \mathbf{Y}]$. As the width of CIs is proportional to the square root of the variance parameters, the posterior CI also gets narrower, which results in poor coverage percentage in the cross validation. On the other hand, the "signal–noise ratio" $\hat{\sigma}_H^2/(\hat{\sigma}_H^2 + \hat{\sigma}_D^2)$ stays almost the same at different Rs. Recall from Section 4.2.2 that this ratio is a key factor in deciding the corrected path, the stable $\hat{\sigma}_H^2/(\hat{\sigma}_H^2 + \hat{\sigma}_D^2)$ ensures that the posterior mean stays almost unchanged and therefore the RMSE's in the cross-validation are nearly equal among the different thinning rates.

We discuss the cause for this seemingly surprising phenomenon in Section 8.2.1. Section 8.2.2 further illustrates the changes in the variance parameter estimates with other data sets and GP models. The implication of this phenomenon and our proposals to alleviate it and fix the credible intervals are discussed in Section 8.2.3 and 8.2.4.

Table 8.1: Posterior modes of $\boldsymbol{\phi}$ $(\hat{\sigma}_H^2, \hat{\sigma}_D^2)$ from $[\boldsymbol{\phi}|\mathbf{X}_R \cup \mathbf{X}_G, \mathbf{Y}]$ and cross validation performance under different rate of thinning for Trip 2 Northing data set. The cross–validation RMSE and coverage rate of the 95% CI (in brackets) from both LOOCV and L5OCV are reported.

Rate	$\hat{\sigma}_H^2$	$\hat{\sigma}_D^2$	$\frac{\hat{\sigma}_H^2}{\hat{\sigma}_H^2+\hat{\sigma}_D^2}$	L5OCV RMSE (Coverage)	LOOCV RMSE (Coverage)
∞	1.03	1.23	0.45	3.06 (93.0%)	0.85 (100%)
1200	0.41	0.45	0.47	2.99 (82.8%)	0.86 (95.3%)
600	0.25	0.27	0.48	2.99 (78.9%)	0.86 (93.8%)
300	0.14	0.15	0.48	3.01 (63.3%)	0.88 (85.9%)
120	0.061	0.066	0.48	3.08 (53.9%)	0.92 (73.4%)
60	0.032	0.034	0.48	3.17 (41.4%)	0.99 (61.7%)

8.2.1 Why the variance parameters decrease

In our simulation studies in Section 4.4.1, we found that the posterior modes of σ_H^2, σ_D^2 are almost the same no matter whether $[\boldsymbol{\phi}|\mathbf{X}_G, \mathbf{Y}]$ or $[\boldsymbol{\phi}|\mathbf{X}, \mathbf{Y}]$ is used. The decrease of variance parameters only happens in the real data set and thus it is probably caused by model misspecification. We provide an explanation for the Brownian Motion process, i.e., decrease in σ_D^2, but similar arguments apply to the σ_H^2 of the Brownian Bridge .

Given a sequence $x_0, x_1, x_2, \ldots, x_n$, observed at equally spaced time points $t_0 = 0, t_1 = \delta, t_2 = 2\delta, \ldots, t_n = n\delta$, we assume that $x_0, x_1, x_2, \ldots, x_n$ are from a zero-mean Brownian Motion process with variance parameter σ_D^2. The posterior mode of $[\sigma_D^2|x_0, x_1, \ldots, x_n]$ under a uniform prior $\log([\sigma_D^2]) \propto 1$ equals the maximum likelihood estimate (MLE) of σ_D^2 on the log scale. Its MLE can be computed in closed form,

$$\hat{\sigma}_D^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - x_{i-1})^2}{\delta}, \tag{8.2}$$

because $(x_i - x_{i-1}) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_D^2\delta)$. When x_i are really from a Brownian Motion, this estimate is "correct", i.e., unbiased and asymptotically consistent. In our simulation studies, $\mathbf{X}$ or its subset $\mathbf{X}_G$ are both from the correct model. When both $\mathbf{X}$ and $\mathbf{X}_G$ have a fairly large n, the estimates of σ_D^2 should be similar, i.e., the probability

that they have a big difference is small. This explains why we do not observe the dramatic change in the simulation study.

However, we have little idea of the behavior of $\hat{\sigma}_D^2$ when the model is misspecified. Now, let's consider an extreme scenario that the x_is are actually from a smooth function $x_i = s(t_i)$, e.g., $s(x) = \sin(x)$. Equation (8.2) is then roughly

$$\hat{\sigma}_D^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - x_{i-1})^2}{\delta} = \frac{1}{n}\sum_{i=1}^{n}\frac{(s(t_i) - s(t_{i-1}))^2}{\delta} \approx \frac{1}{n}\sum_{i=1}^{n}(s'(t_{i-1}))^2\delta.$$

If we assume $|s'(t)|^2 \approx C$ for all t[1], the estimate of σ_D^2 is approximately $C\delta$. The smaller δ is, the smaller $\hat{\sigma}^2$. Under this misspecified model, when δ is large, the data appear to be wiggly for the model, which results in a big σ_D^2 estimate. But it appears to be smoother with smaller δ, which leads to a smaller $\hat{\sigma}^2$.

Our real data are not such a smooth function, but it certainly has a smooth component as seen in Figure 4.6. When more of $\mathbf{X}$ from $\mathbf{X}$ is included in the inference, the data used (random bias part in DR $\boldsymbol{\xi}$) appear to be smoother and resemble a Brownian Motion with smaller variance. This explains the substantial decrease of the variance parameters in Table 8.1.

8.2.2 Similar phenomenon in other data sets and models

The BM-GOU models developed in Chapter 6 also experience such a decrease in the variance parameters σ, σ_D^2 when more of $\mathbf{X}$ is used in $[\boldsymbol{\phi}|\mathbf{X}_R \cup \mathbf{X}_G, \mathbf{Y}]$, but the estimates of the B-spline coefficients $\boldsymbol{\gamma}^{(A)}$ and $\boldsymbol{\gamma}^{(a)}$ stay almost the same. A similar phenomenon can also be observed even when we work with a "smoother" process, such as the CRAWL and Matérn process. The CRAWL works with only the GPS observation, so we create a similar situation by thinning them at different rates (i.e., thinning rate equals 2 means every other observation). For the Matérn model, we assume the GPS observations are from a Matérn process with smooth parameter $\nu = 1.5$ plus observation noises with a known variance $\sigma_G^2 = 0.0625$, which is a similar setting to that of our Bayesian Melding models. The estimates of the variance parameters in the CRAWL model and Matérn process at different thinning rates for the Trip 2 (Northing) data set are summarized in Table 8.2, which

[1] The square of the first derivative ranges in a range around C.

shows that the variance parameters for these two models also decrease with the decrease of subset rate in general. When compared to Table 8.1, the changes in the variance parameters are relatively moderate and they do not decrease linearly as in Table 8.1.

Table 8.2: The estimates of variance parameters in CRAWL and Matérn model with different subset rates for the Trip 2 (Northing) GPS data set. Thinning rate = 1 gives the original data set.

Thinning Rate	4	3	2	1
$\hat{\sigma}^2$ in CRAWL	10137.5	9759.5	9172.8	542.3
$\hat{\sigma}^2$ in Matérn	979.6	1079.1	1042.3	959.4

Besides the animal tracking data, we also performed similar analysis on the motorcycle data set, which was first published in Silverman (1985). It contains a series of measurements of head acceleration in a simulated motorcycle accident, used to test crash helmets. These data are plotted in Figure 8.1. As with our animal tracks, the data also have a smooth component. The variance parameter estimates using the model of Brownian Motion or Matérn process with $\nu = 1.5$ are summarized in Table 8.3. Notice that we assume there is no observation noise, i.e., $\sigma_G^2 = 0$. Table 8.3 shows that the $\hat{\sigma}^2$s change dramatically in both the Brownian Motion and Matérn model for the motorcycle data set. Unlike the seal tracking data set, they are not changing monotonically with the thinning rate. This is probably caused by the fact that a lot of time points are quite close to each other in this data set and we choose not to consider the observation noise in this illustration. We only aim to demonstrate changes in the variance parameter estimates and will not further study this data set.

Table 8.3: Variance parameter estimates of the Brownian Motion and Matérn process for the motorcycle data set.

Thinning Rate	4	3	2	1
$\hat{\sigma}^2$ for Brownian Motion	609.6	495.1	991.1	2643.4
$\hat{\sigma}^2$ for Matérn Process	2152.3	1895.6	2160.8	2401.4

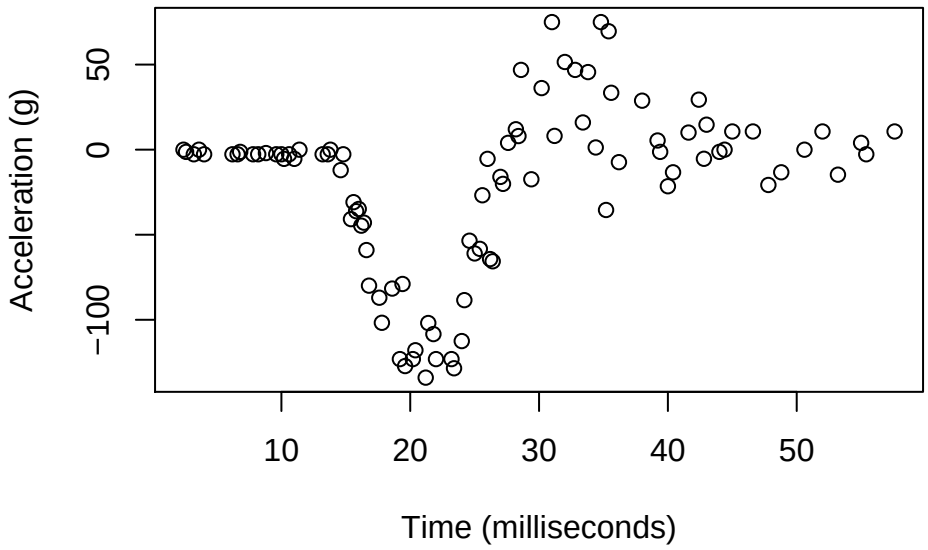


Figure 8.1: The motorcycle data set.

8.2.3 Ideas to stabilize the variance parameters

From the above tables, it is clear that the sampling rate can have a substantial impact on the estimates of variance parameters, which then affect the credible intervals for the predictions. The problem is caused by the fact that the GP model does not fit the smooth component in the data.

The smoother GP's, such as the CRAWL and Matérn can help to alleviate the dramatic change to some degree, but they do not completely remove this effect as shown in Tables 8.2 and 8.3. These smoother GP's can also involve strong parametric assumptions on the data, which may not work well for real data. For example, we have shown that the CRAWL (integrated Ornstein Ulhenbeck process) has worse predictive performance than linear interpolation in Section 4.5.4. The Matérn process is shown to have inconsistent parameter estimates under certain designs (Zhang, 2004) and we find it is numerically difficult to calculate its estimates. Those smoother GPs, such as Matérn with $\nu > 0.5$, do not have a Markovian structure and require extra effort to scale it for big data as reviewed in Section 3.1 (Lindgren et al., 2011).

Another idea is to model the smooth component with a flexible mean structure

parametrically, such as with a spline or polynomials (see e.g., Buderman et al., 2015). The limitation of this idea is that we need many parameters to ensure that the splines or polynomials are flexible enough, which makes the inference difficult given the limited data. Also, it is tricky to ensure the quality of extrapolation or interpolation in periods where we do not have observations, as the parametric structure of the splines/polynomials will have a strong effect on the predictions. We have made two attempts in this direction for the seal tracking study: the polynomials for the bias component of DR and the B-splines in the GOU process. Neither of them is successful in stabilizing the variance parameter at different sampling frequencies, but the GOU achieves better predictive performance.

In particular, the nature of the seal tracking data set makes it difficult for both ideas to succeed. Recall that the GPS observations are only available at a sparse set of time points. Even if the true path were from a smooth GP, like CRAWL or Matérn, the dependence would be very weak if two GPS observations were far apart. So a wiggly process such as the Brownian Bridge/Motion may fit the data adequately. The sparsity of the GPS observations also makes it difficult to fit these mean curves or high–order dependence models. The DR path has high resolution, but we cannot distinguish the true path from the noise except at those time points when GPS sitings are available. This is probably the main reason that the high order polynomials do not improve the performance as in Section 4.5.2.

We also point out that the flexible mean and smooth covariance can be confounded. For example, Watson (1984) and Silverman (1985), have discussed a certain equivalence between the regression spline and kriging. Fan et al. (2014) discussed certain equivalence between the higher order stochastic differential equations and the Chebyshev splines. We should be aware of this issue when trying both ideas simultaneously.

8.2.4 The variance parameters and the CI

If we are only interested in the mean of the predictions, the variance parameter actually does not matter, as shown in CV-RMSE in Table 8.1. However, the variance parameters are important in the prediction uncertainty. We cannot fully explain why conditioning on $\mathbf{X}_G$ leads to "correct" coverage percentages in the L5OCV

regardless of the model (i.e., BM-GOU, BM-BB and DS), but not in the LOOCV. A similar question is why subset rate $r = 1200$ leads to almost correct coverage percentage in the LOOCV, but not the L5OCV. There are two aspects of this question: which cross–validation scheme we should use and which subset rate can lead to a "correct" variance parameter in the cross–validation.

The reason that we chose L5OCV over LOOCV when evaluating our model performance was discussed in Section 4.5. In short, the left-out observation is still very close to the nearby remaining observations in the LOOCV, such that the dependence from the nearby remaining observations helps to determine an interval that covers the left-out observation. The L5OCV reduces such dependence and its intervals are mainly decided by our models. We can confirm this point from Table 8.1 as the LOOCV always has higher coverage percentage than the L5OCV regardless of the subset rates.

The second part of this question is more complicated as it involves both the behavior of our model and the true path. Ideally, we desire the model to capture the data's uncertainty and automatically adapt to different data and prediction requirements, i.e., same correct coverage percentage regardless of the subset rate or which cross validation to use. Stabilizing the variance parameter as discussed in the above subsection can be a first step to tackle this problem.

Given this coverage issue, we cannot claim our models result in CIs with correct coverage percentage for all the other tracks with GPS and DR paths in any cross–validation scheme. First, we stress that the cross–validation scheme, i.e., how much to leave out, should be decided by the scientific objective of the study, i.e., how long a time gap the path is expected to predict. Another way to decide the cross–validation scheme is to make the left-out data nearly independent from the training data (see e.g., Arlot et al., 2010). After the CV is decided, the amount of uncertainty in our model can be controlled via the priors in our framework. Informative priors can force the variance parameters into values that can lead to correct coverage percentages. The parameters in the prior may be chosen by a combination of prior knowledge and cross–validation.

This "informative prior + cross–validation selection for priors" scheme of using our Bayesian Melding model is not ideal, but it should work for practical purposes. As said by George Box: "All models are wrong, but some are useful." The seals or

basketball players do not move according to any mathematical formulas. The models developed in this project are our best effort to express these tracks statistically and use them for interpretation or prediction via moderate computational power.

8.3 Future work

Besides the issue of variance parameters discussed above, there are many aspects of this thesis that can be further developed. The possible rescue of CHGP has been discussed by the end of Chapter 5. Similarly, the future development of the PF model for NBA tracking data has been discussed in Section 7.1. The rest of this section covers the future directions for our BM framework and PF approach.

8.3.1 Future developments for Bayesian melding

Our work has demonstrated the value of using the Bayesian data fusion techniques to combine observations from different sources for tracking objects. We plan to adapt the ideas developed in this paper to track the progress of infectious disease. We will use the Google flu trends data as a biased but high resolution source of observations and Center for Disease Control (CDC) reports as an accurate but sparse set of observations. The bias and failure to predict some disease epidemics by Google flu have been studied in Hooten et al. (2010); Lazer et al. (2014b,a) and the references within. The media also used it as a serious warning about the use of "big data". As noted by Salzberg (2014) "Big data can be great, but not when it is bad data." While agreeing with this comment in principle, we point out that "bad big data", such as the Google flu trend or the DR path in our paper, can be good again in cases when it can be combined and corrected by good data. So for example, Dugas et al. (2013) and Lazer et al. (2014b) used various regression models to combine the Google flu and CDC reports and achieved better predictions for infectious disease. Our BM and downscaling approaches also successfully combine the DR path with GPS observations.

Another interesting direction is to use a SDE parameterized by PF as the prior process in our BM framework. The implementation of this idea is straightforward as the Euler approximation can express the SDE in a DLM and then the Kalman smoother can be used. The main challenge might be how to find a parsimonious

parameterization for the PF in the spatio–temporal domain. In our current work, we considered coefficients that vary in the time domain in Chapter 6 and vary in the spatial domain in Chapter 7. A PF reflecting both changes in the spatial and temporal domains is key to its success in interpreting the tracks, such as those from the fur seals. When working with a mixture of PFs in BM, we may need to design new approximation techniques to avoid obtaining Monte Carlo samples on the hidden classes. A starting point is marginalizing the latent classes in the PF and using the marginalized process in the BM framework.

8.3.2 Approximation to non-linear SDE

In the above discussion of the PF, the nonlinear PFs were all approximated by the local constant Euler approximation. Because our tracking data were observed at high resolution, the Euler approximation is accurate (Iacus, 2009). Yet, if a better approximation is desired, we can consider other approximation techniques, such as Ozaki (1992); Archambeau et al. (2007); Archambeau and Opper (2010); Vrettas et al. (2015), which have been briefly reviewed in Section 3.2. It is beyond the scope of this thesis to introduce their formulation and compare their differences. In particular, we find that the Euler and Ozaki (Ozaki, 1992) approximations can both be derived from the Kolmogorov backward equation (see e.g., Arnold, 1974). An interesting idea is to derive similar approximations from the Kolmogorov forward equation or combine them in some way.

Bibliography

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342. → pages 79

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007). Gaussian process approximations of stochastic differential equations. *Gaussian Processes in Practice*, 1:1–16. → pages 25, 115

Archambeau, C. and Opper, M. (2010). Approximate inference for continuous-time Markov processes. *Bayesian Time Series Models*, pages 125–140. → pages 25, 115

Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79. → pages 113

Arnold, L. (1974). *Stochastic Differential Equations*. Wiley-Interscience. → pages 24, 25, 78, 79, 81, 82, 83, 115

Ba, S. and Joseph, V. R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, 6(4):1838–1860. → pages 20

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848. → pages 13, 15, 66

Bates, D. and Maechler, M. (2016). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-7.1. → pages 12, 84

Battaile, B. (2014). *TrackReconstruction: Reconstruct animal tracks from magnetometer, accelerometer, depth and optional speed data.* R package version 1.1. → pages 9

Battaile, B. C., Nordstrom, C. A., Liebsch, N., and Trites, A. W. (2015). Foraging a new trail with northern fur seals (Callorhinus ursinus): Lactating seals from islands with contrasting population dynamics have different foraging strategies, and forage at scales previously unrecognized by GPS interpolated dive data. *Marine Mammal Science*, 31(4):1494–1520. → pages 5, 33

Benoit-Bird, K. J., Battaile, B. C., Heppell, S. A., Hoover, B., Irons, D., Jones, N., Kuletz, K. J., Nordstrom, C. A., Paredes, R., Suryan, R. M., Waluk, C. M., and Trites, A. W. (2013a). Prey patch patterns predict habitat use by top marine predators with diverse foraging strategies. *PloS ONE*, 8(1):e53348. → pages 4, 8

Benoit-Bird, K. J., Battaile, B. C., Nordstrom, C. A., and Trites, A. W. (2013b). Foraging behavior of northern fur seals closely matches the hierarchical patch scales of prey. *Marine Ecology Progress Series*, 479:283–302. → pages 4, 36

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122. → pages 71

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(2):176–197. → pages 31, 32, 34, 44

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons. → pages 20

Block, B. A., Jonsen, I. D., Jorgensen, S. J., Winship, A. J., Shaffer, S. A., Bograd, S. J., Hazen, E. L., Foley, D. G., Breed, G. A., Harrison, A.-L., Ganong, J. E., Swithenbank, A., Castleton, M., Dewar, H., Mate, B. R., Shillinger, G. L., Schaefer, K. M., Benson, S. R., Weise, M. J., Henry, R. W., and Costa, D. P. (2011). Tracking apex marine predator movements in a dynamic ocean. *Nature*, 475:86–90. → pages 4

Boehme, L., Kovacs, K., and Lydersen, C. (2010). Biologging in the global ocean observing system. *Proceedings of OceanObs 09: Sustained Ocean Observations and Information for Society*, 2(September):21–25. → pages 4

Boehme, L., Meredith, M. P., Thorpe, S. E., Biuw, M., and Fedak, M. (2008). Antarctic Circumpolar Current frontal system in the South Atlantic: Monitoring using merged Argo and animal-borne sensor data. *Journal of Geophysical Research: Oceans*, 113(C9). → pages 4

Bornn, L., Shaddick, G., and Zidek, J. V. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, 107(497):281–289. → pages 19

Brillinger, D., Preisler, H., Ager, A., and Wisdom, M. (2004). Stochastic differential equations in the analysis of wildlife motion. → pages 93

Brillinger, D. R. et al. (2007). A potential function approach to the flow of play in soccer. *Journal of Quantitative Analysis in Sports*, 3(1). → pages 93, 101

Brillinger, D. R. and Stewart, B. S. (2010). Stochastic modeling of particle movement with application to marine biology and oceanography. *Journal of Statistical Planning and Inference*, 140(12):3597–3607. → pages 93

Bryant, E. (2007). 2D location accuracy statistics for Fastloc RCores running firmware versions 2.2 & 2.3. *Wildtrack Telemetry Systems Ltd*. → pages 5, 38

Buderman, F. E., Hooten, M. B., Ivan, J. S., and Shenk, T. M. (2015). A functional model for characterizing long-distance movement behaviour. *Methods in Ecology and Evolution*, (7):264–273. → pages 30, 112

Caragea, P. C. and Smith, R. L. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98(7):1417–1440. → pages 18

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87. → pages 137

Cervone, D., Bornn, L., and Goldsberry, K. (2016). NBA court realty. *SLOAN Sports Analytics Conference, Boston*. → pages 94

Cervone, D., D'Amour, A., Bornn, L., and Goldsberry, K. (2014). POINTWISE: predicting points and valuing decisions in real time with NBA optical tracking data. *Proceedings MIT Sloan Sports Analytics*. → pages 1, 94

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *The Journal of Finance*, 47(3):1209–1227. → pages 78

Chipman, H. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948. → pages 13, 75

Chu, T., Wang, H., and Zhu, J. (2014). On semiparametric inference of geostatistical models via local Karhunen–Loève expansion. *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology), pages 817–832. → pages 19

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226. → pages 15

Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-temporal Data*. John Wiley & Sons. → pages 11, 14

Damian, D., Sampson, P. D., and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, 12(2):161–178. → pages 19

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812. → pages 13, 15, 16, 17, 18, 65, 66

Davis, T. A. (2006). *Direct Methods for Sparse Linear Systems*, volume 2. SIAM. → pages 84

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978). *A Practical Guide to Splines*, volume 27. Springer-Verlag New York. → pages 30, 95

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436. → pages 79

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38. → pages 102

Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., and Rothman, R. E. (2013). Influenza forecasting with Google flu trends. *PLoS ONE*, 8(2):e56176. → pages 114

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google flu trends data and a state-space SEIR model. *Journal of the American Statistical Association*, 107(500):1410–1426. → pages 33

Eidsvik, J. and Shaby, B. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315. → pages 18, 66

Elkaim, G. H., Decker, E. B., Oliver, G., and Wright, B. (2006). Marine Mammal Marker (MAMMARK) dead reckoning sensor for in-situ environmental monitoring. *Proceedings of the ION/IEEE PLANS*. → pages 6

Fan, R., Zhu, B., and Wang, Y. (2014). Stochastic dynamic models and Chebyshev splines. *Canadian Journal of Statistics*, 42(4):610–634. → pages 112

Fleming, C. H., Fagan, W. F., Mueller, T., Olson, K. A., Leimgruber, P., and Calabrese, J. M. (2016). Estimating where and how animals travel: An optimal framework for path reconstruction from autocorrelated tracking data. *Ecology*. → pages 6

Foley, K. M. and Fuentes, M. (2008). A statistical framework to combine multivariate spatial data and physical models for hurricane surface wind prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, 13(1):37–59. → pages 32

Franks, A., Miller, A., Bornn, L., Goldsberry, K., et al. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9(1):94–121. → pages 94

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer. → pages 139

Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45. → pages 31, 32, 35

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523. → pages 12

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014. → pages 1

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130. → pages 13, 75

Grewal, M. S., Weill, L. R., and Andrews, A. P. (2007). *Global Positioning Systems, Inertial Navigation, and Integration*. John Wiley & Sons. → pages 6

Grün, B., Leisch, F., et al. (2008). Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35. → pages 102

Gurarie, E., Andrews, R. D., and Laidre, K. L. (2009). A novel method for identifying behavioural changes in animal movement data. *Ecology Letters*, 12(5):395–408. → pages 68

Hazel, J. (2009). Evaluation of fast-acquisition GPS in stationary tests and fine-scale tracking of green turtles. *Journal of Experimental Marine Biology and Ecology*, 374(1):58–68. → pages 5

Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60. → pages 15, 40

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian Statistics*, 6(1):761–768. → pages 19, 20

Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J. (2012). *Global Positioning System: Theory and Practice*. Springer Science & Business Media. → pages 5

Hooten, M., Johnson, D., McClintock, B., and Morales, J. (2017). *Animal Movement: Statistical Models for Telemetry Data*. Chapman & Hall/CRC Boca Raton, Florida, USA. → pages 93

Hooten, M. B., Anderson, J., and Waller, L. A. (2010). Assessing north american influenza dynamics with a statistical sirs model. *Spatial and Spatio-temporal Epidemiology*, 1(2):177–185. → pages 114

Hooten, M. B., Hanks, E. M., Johnson, D. S., and Alldredge, M. W. (2013). Reconciling resource utilization and resource selection functions. *The Journal of Animal Ecology*, 82(6):1146–54. → pages 106

Hooten, M. B. and Johnson, D. S. (2017). Basis function models for animal movement. *Journal of the American Statistical Association*, (just-accepted). → pages 6, 20

Horne, J. S., Garton, E., Krone, S., and Lewis, J. (2007). Analyzing animal movements using Brownian bridges. *Ecology*, 88(9):2354–2363. → pages 36

Huang, H., Cressie, N., and Gabrosek, J. (2002). Fast, resolution-consistent spatial prediction of global processes from satellite data. *Journal of Computational and Graphical Statistics*, 11(1):63–88. → pages 14

Humphries, N. E., Queiroz, N., Dyer, J. R. M., Pade, N. G., Musyl, M. K., Schaefer, K. M., Fuller, D. W., Brunnschweiler, J. M., Doyle, T. K., Houghton, J. D. R., Hays, G. C., Jones, C. S., Noble, L. R., Wearmouth, V. J., Southall, E. J., and Sims, D. W. (2010). Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature*, 465(7301):1066–9. → pages 36

Iacus, S. M. (2009). *Simulation and Inference for Stochastic Differential Equations: with R Examples*. Springer Science & Business Media. → pages 23, 25, 115

Johnson, D., London, J., Lea, M.-a., and Durban, J. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology*, 89(5):1208–1215. → pages 6, 54

Johnson, M. P. and Tyack, P. L. (2003). A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE Journal of Oceanic Engineering*, 28(1):3–12. → pages 5

Jonsen, I., Flemming, J., and Myers, R. (2005). Robust state-space modeling of animal movement data. *Ecology*, 86(11):2874–2880. → pages 6

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45. → pages 28

Katzfuss, M. (2016). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, (just-accepted). → pages 66

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555. → pages 12

Kranstauber, B., Kays, R., Lapoint, S. D., Wikelski, M., and Safi, K. (2012). A dynamic Brownian bridge movement model to estimate utilization distributions for heterogeneous animal movement. *The Journal of Animal Ecology*, 81(4):738–46. → pages 36, 68

Kranstauber, B., Safi, K., and Bartumeus, F. (2014). Bivariate Gaussian bridges: directional factorization of diffusion in Brownian bridge models. *Movement Ecology*, 2(1):5. → pages 36, 68

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford, UK. → pages 132

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014a). Google flu trends still appears sick: An evaluation of the 2013-2014 flu season. *SSRN Electronic Journal*, pages 1–11. → pages 114

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014b). The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–1205. → pages 1, 114

Le, N. D. and Zidek, J. V. (2006). *Statistical Analysis of Environmental Space-time Processes*. Springer. → pages 11, 63

Leis, J. W. (2011). *Digital Signal Processing Using MATLAB for Students and Researchers*. John Wiley & Sons. → pages 63

Lemos, R. T. and Sansó, B. (2012). Conditionally linear models for non-homogeneous spatial random fields. *Statistical Methodology*, 9(1-2):275–284. → pages 19

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498. → pages 13, 18, 19, 20, 34, 111

Liu, Y. (2014). *BayesianAnimalTracker: Bayesian Melding of GPS and DR Path for Animal Tracking*. R package version 1.2. → pages 42

Liu, Y., Battaile, B. C., Zidek, J. V., and Trites, A. W. (2015). Bias correction and uncertainty characterization of Dead-Reckoned paths of marine mammals. *Animal Biotelmetry*, 3(51). → pages 7, 42, 49, 50, 62

Liu, Y., Dinsdale, D., Jun, S.-H., Briercliffe, C., and Bone, J. (2016). Statistical learning of basketball strategy: The potential field approach. *Cascadia Symposium on Statistics in Sports, Vancouver*. → pages 1, 34

Liu, Z., Le, N. D., and Zidek, J. V. (2011). An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics*, 22(3):340–353. → pages 32, 137

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67:68–83. → pages 20, 44

McClintock, B. T., Johnson, D. S., Hooten, M. B., Ver Hoef, J. M., and Morales, J. M. (2014). When to be discrete: the importance of time formulation in understanding animal movement. *Movement ecology*, 2(1):21. → pages 5, 62

McIntyre, A., Brooks, J., Guttag, J., and Wiens, J. (2016). Recognizing and analyzing ball screen defense in the nba. *SLOAN Sports Analytics Conference, Boston*. → pages 94

Miller, A., Bornn, L., Adams, R., and Goldsberry, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. *arXiv preprint arXiv:1401.0942*. → pages 1

Nordstrom, C. A., Benoit-Bird, K. J., Battaile, B. C., and Trites, A. W. (2013). Northern fur seals augment ship-derived ocean temperatures with higher temporal and spatial resolution data in the eastern Bering Sea. *Deep Sea Research Part II*, 94:257–273. → pages 4, 8

Nychka, D. W., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multi-resolution Gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics*, 24(2):579–599. → pages 11, 13, 15, 16, 19, 66

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138. → pages 70, 71

Ozaki, T. (1992). A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: A local linearization approach. *Statistica Sinica*, 2(1):113–135. → pages 25, 79, 115

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506. → pages 19, 20

Park, C., Huang, J., and Ding, Y. (2011). Domain decomposition approach for fast Gaussian process regression of large spatial data sets. *The Journal of Machine Learning Research*, 12:1697–1728. → pages 13

Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer. → pages 26, 28, 84

Pissanetzky, S. (1984). *Sparse Matrix Technology-electronic Edition*. Academic Press. → pages 12

Pozdnyakov, V., Meyer, T., Wang, Y.-B., and Yan, J. (2014). On modeling animal movements using Brownian motion with measurement error. *Ecology*, 95(2):247–253. → pages 36

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. → pages 95

Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer Science & Business Media. → pages 30

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2014). *fda: Functional Data Analysis*. R package version 2.4.4. → pages 81

Rogers, L. C. G. and Williams, D. (2000). *Diffusions, Markov processes and Martingales: Volume 2, Itô Calculus*, volume 2. Cambridge university press. → pages 79

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC Press. → pages 11, 13, 18, 83

Rue, H. v., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392. → pages 13, 20, 21, 34, 39, 137, 138

Rundel, C. W., Schliep, E. M., Gelfand, A. E., and Holland, D. M. (2015). A data fusion approach for spatial analysis of speciated PM2.5 across time. *Environmetrics*, 26(8):515–525. → pages 32

Russell, J. C., Hanks, E. M., Haran, M., and Hughes, D. P. (2016). A spatially-varying stochastic differential equation model for animal movement. *arXiv preprint arXiv:1603.07630*. → pages 93

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423. → pages 35

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2010). Fusing point and areal level space–time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(1):77–103. → pages 32

Salzberg, S. (2014). Why google flu is a failure. *Forbes. com [online]*, pages 03–24. → pages 114

Sampson, P. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Associationan*, 87(417):108–119. → pages 19

Särkkä, S., Solin, A., and Hartikainen, J. (2013). Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61. → pages 13

Sawyer, H., Kauffman, M. J., Nielson, R. M., and Horne, J. S. (2009). Identifying and prioritizing ungulate migration routes for landscape-level conservation. *Ecological Applications*, 19(8):2016–2025. → pages 36

Scharf, H. R., Hooten, M. B., Fosdick, B. K., Johnson, D. S., London, J. M., Durban, J. W., et al. (2017). Dynamic social networks based on movement. *The Annals of Applied Statistics*, 10(4):2182–2202. → pages 100

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758. → pages 19

Shaddick, G. and Zidek, J. V. (2015). *Spatio-temporal Methods in Environmental Epidemiology*. CRC Press. → pages 11

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1):1–52. → pages 110, 112

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493. → pages 53

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. → pages 53

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19. → pages 16

Stein, M. L., Chi, Z., and Welty, L. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296. → pages 18, 66

Stirzaker, G. G. D. and Grimmett, D. (2001). *Probability and Random Processes*. Oxford Science Publications. → pages 40, 131

Tzeng, S., Huang, H.-C., and Cressie, N. (2005). A fast, optimal spatial-prediction method for massive datasets. *Journal of the American Statistical Association*, 100(472):1343–1357. → pages 14

Vecchia, A. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 50(2):297–312. → pages 18

Vrettas, M. D., Opper, M., and Cornford, D. (2015). Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. *Physical Review E*, 91(1):012148. → pages 25, 79, 115

Wahba, G. (1965). A least squares estimate of satellite attitude. *SIAM Review*, 7(3):409–409. → pages 7

Wang, K.-C. and Zemel, R. (2016). Classifying NBA offensive plays using neural networks. *SLOAN Sports Analytics Conference, Boston*. → pages 94

Watson, G. (1984). Smoothing and interpolation by kriging and with splines. *Journal of the International Association for Mathematical Geology*, 16(6):601–615. → pages 112

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. → pages 11

Wilson, R. P., R. and Wilson, M. P. (1988). Dead reckoning–a new technique for determining penguin movements at sea. *Meeresforschung–Reports on Marine Research*, 32:155–158. → pages 5, 6, 9

Wilson, R. P., Liebsch, N., Davies, I. M., Quintana, F., Weimerskirch, H., Storch, S., Lucke, K., Siebert, U., Zankl, S., Müller, G., Zimmer, I., Scolaro, A., Campagna, C., Plötz, J., Bornemann, H., Teilmann, J., and McMahon, C. R. (2007). All at sea with animal tracks; methodological and analytical solutions for the resolution of movement. *Deep Sea Research Part II*, 54(3-4):193–210. → pages 1, 4, 5, 6, 7, 8, 9

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261. → pages 111

Zhu, B., Song, P. X.-K., and Taylor, J. M. (2011a). Stochastic functional data analysis: A diffusion model-based approach. *Biometrics*, 67(4):1295–1304. → pages 78

Zhu, B., Taylor, J. M., and Song, P. X.-K. (2011b). Semiparametric stochastic modeling of the rate function in longitudinal studies. *Journal of the American Statistical Association*, 106(496):1485–1495. → pages 78

Zidek, J. V., Le, N. D., and Liu, Z. (2012). Combining data and simulated data for spacetime fields: application to ozone. *Environmental and Ecological Statistics*, 19(1):37–56. → pages 31, 32, 34, 44

Appendix A

Inferential Methods for Bayesian Melding

This section includes the detailed derivations of the inferential methods needed for BM.

A.1 Explicit form of $\langle \beta, \eta | \mathbf{X}, \mathbf{Y} \rangle$

Following the model and notation in (4.2) to (4.7), it is easy to find

$$\begin{aligned}\langle \boldsymbol{\beta}, \boldsymbol{\eta} | \mathbf{X}, \mathbf{Y} \rangle \propto & \langle \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{X}, \mathbf{Y} \rangle = \langle \boldsymbol{\eta} \rangle \langle \mathbf{Y} | \boldsymbol{\eta} \rangle \langle \mathbf{X} | \boldsymbol{\beta}, \boldsymbol{\eta} \rangle \\ \propto & |\mathbf{R}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\eta} - \mathbf{f})^T \mathbf{R}^{-1} (\boldsymbol{\eta} - \mathbf{f}) \right\} \\ & \times |\mathbf{D}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{Y} - \mathbf{G}\boldsymbol{\eta})^T \mathbf{D}^{-1} (\mathbf{Y} - \mathbf{G}\boldsymbol{\eta}) \right\} \\ & \times |\mathbf{C}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{X} - \mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\eta})^T \mathbf{C}^{-1} (\mathbf{X} - \mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\eta}) \right\},\end{aligned} \tag{A.1}$$

where the following notation is used in addition to those from (4.2) to (4.7):

- $\mathbf{f} = (f(1), f(2), \ldots, f(T-1))^T$, is a $T-1$ vector as the prior mean of the random part in $\boldsymbol{\eta}$.
- $\mathbf{D} = \sigma_G^2 \mathbf{I}_{K-1}$ is the covariance matrix of $\mathbf{Y}$ conditional on $\boldsymbol{\eta}$, where $\mathbf{I}_m$ stands for the identity matrix of dimension m.

- $\mathbf{Z}$ is the design matrix for $\mathbf{h}$ as in (4.5).
- $\mathbf{G}$ is a $(K-1)\times(T-1)$ matrix, with

$$G_{k,j}=\begin{cases}1, & j=t_k-1\\ 0, & \text{Otherwise}\end{cases},$$

  for $k=1,2,\ldots,K-1$. Notice that the first time point is removed.
- $\mathbf{R}=\mathbf{R}(1:(T-1),1:(T-1))$ is the $(T-1)\times(T-1)$ covariance matrix for the Brownian Bridge.
- $\mathbf{C}=\mathbf{C}(1:T,1:T)$ is the $T\times T$ covariance matrix for the Brownian Motion error term $\boldsymbol{\xi}$.

To further simplify (A.1), the following notation is needed:

- $\boldsymbol{\zeta}=(\boldsymbol{\beta}^T,\boldsymbol{\eta}^T)^T$, which is the joint vector of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ of length $Q+T-1$.
- $\mathbf{U}$ is a $T\times(Q+T-1)$ matrix with $\mathbf{U}(1:T,1:Q)=\mathbf{Z}$, $\mathbf{U}(1:(T-1),(Q+1):(Q+T-1))=I_{T-1}$ and $\mathbf{U}(T,(Q+1):(Q+T-1))=0$, which maps $\mathbf{U}\boldsymbol{\zeta}=\mathbf{h}+\boldsymbol{\eta}$.
- $\mathbf{V}$ is a $(T-1)\times(Q+T-1)$ matrix with $\mathbf{V}(1:(T-1),1:Q)=\mathbf{0}$ and $\mathbf{V}(1:(T-1),(Q+1):(Q+T-1))=\mathbf{I}_{T-1}$, such that $\mathbf{V}\boldsymbol{\zeta}=\boldsymbol{\eta}$.
- $\mathbf{W}$ is a $(K-1)\times(Q+T-1)$ matrix with $\mathbf{W}(1:(K-1),1:Q)=0$ and $\mathbf{W}(1:(K-1),(Q+1):(Q+T-1))=\mathbf{G}$, such that $\mathbf{W}\boldsymbol{\zeta}=\mathbf{G}\boldsymbol{\eta}$.

Some algebra simplifies (A.1) and yields

$$\begin{aligned}\langle\boldsymbol{\zeta}|\mathbf{X},\mathbf{Y}\rangle \propto & |\mathbf{R}|^{-1/2}|\mathbf{D}|^{-1/2}|\mathbf{C}|^{-1/2}\times \\ & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\zeta}-\mathbf{M}_1^{-1}\mathbf{M}_2)^T\mathbf{M}_1(\boldsymbol{\zeta}-\mathbf{M}_1^{-1}\mathbf{M}_2)+M_3-\mathbf{M}_2^T\mathbf{M}_1^{-1}\mathbf{M}_2\right]\right\}, \\ \propto & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\zeta}-\mathbf{M}_1^{-1}\mathbf{M}_2)^T\mathbf{M}_1(\boldsymbol{\zeta}-\mathbf{M}_1^{-1}\mathbf{M}_2)\right]\right\}\end{aligned} \tag{A.2}$$

where

$$\begin{aligned}
\mathbf{M}_1 =& \mathbf{V}^T\mathbf{R}^{-1}\mathbf{V} + \mathbf{W}^T\mathbf{D}^{-1}\mathbf{W} + \mathbf{U}^T\mathbf{C}^{-1}\mathbf{U},\\
\mathbf{M}_2 =& \mathbf{V}^T\mathbf{R}^{-1}\mathbf{f} + \mathbf{W}^T\mathbf{D}^{-1}\mathbf{Y} + \mathbf{U}^T\mathbf{C}^{-1}\mathbf{X},\\
M_3 =& \mathbf{f}^T\mathbf{R}^{-1}\mathbf{f} + \mathbf{Y}^T\mathbf{D}^{-1}\mathbf{Y} + \mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}.
\end{aligned}$$

The posterior distribution of $\boldsymbol{\beta}, \boldsymbol{\eta}$ conditional on $\boldsymbol{\phi}$ is a multivariate Gaussian density. The main computational complexity involved in evaluating this density comes from the calculation of $\mathbf{M}_1^{-1}\mathbf{M}_2$.

A.2 Derivation of (4.11) and (4.12)

It is well known that the Brownian Bridge and Brownian Motion processes are Markovian (Stirzaker and Grimmett, 2001), such that:

$$\begin{aligned}
[\eta(t)|\eta(t-1), \eta(t-2)] =& [\eta(t)|\eta(t-1)]\\
[\xi(t)|\xi(t-1), \xi(t-2)] =& [\xi(t)|\xi(t-1)],
\end{aligned}$$

where $\eta(t)$ is a Brownian Motion process as in our model (4.2) and $\xi(t)$ is a Brownian Motion process as in (4.5). This Markovian property directly suggests the conditional independence property of these two processes, such that:

$$\begin{aligned}
[\eta(t-1), \eta(t+1)|\eta(t)] =& [\eta(t-1)|\eta(t)][\eta(t+1)|\eta(t)], \text{and}\\
[\xi(t-1), \xi(t+1)|\xi(t)] =& [\xi(t-1)|\xi(t)][\xi(t+1)|\xi(t)].
\end{aligned}$$

These properties help us derive (4.11) and (4.12). As an illustration, consider the case where $T = 4$ (Recall that $\eta(0)$ and $\eta(4)$ are fixed, so only $\boldsymbol{\eta}(1,2,3)$ are random), and one GPS observation is available at $t = 2$. The first part of (4.10), $\langle \boldsymbol{\eta}(1 : T \setminus t_{1:K})|\boldsymbol{\beta}, \boldsymbol{\eta}_G, \mathbf{X}, \mathbf{Y}\rangle$ in this situation is

$$\langle \eta(1), \eta(3)|\eta(2), \mathbf{X}(1:3), Y(2), \boldsymbol{\beta}\rangle = \langle \eta(1), \eta(3)|\eta(2), \mathbf{X}(1:3), \boldsymbol{\beta}\rangle, \qquad \text{(A.3)}$$

as $Y(2)$ only depends on $\eta(2)$ and is independent of all the other random variables. For brevity, we omit the $\boldsymbol{\beta}$ term in all the following derivations and let

$X_i = X(i), \eta_i = \eta(i), \xi_i = \xi(i)$. Using the conditional independence property of η and ξ together with certain variable transformations, we can simplify (A.3) into

$$\begin{aligned}
\langle \eta_1, \eta_3 | \eta_2, X_{1:3} \rangle &= \frac{\langle \eta_1, \eta_2, \eta_3, X_1, X_2, X_3 \rangle}{\langle \eta_2, X_1, X_2, X_3 \rangle} = \frac{\langle \eta_1, \eta_2, \eta_3, \xi_1, \xi_2, \xi_3 \rangle}{\langle \eta_2, X_1, X_2, X_3 \rangle} \\
&= \frac{\langle \eta_1, \eta_2, \eta_3 \rangle \langle \xi_1, \xi_2, \xi_3 \rangle}{\langle \eta_2, X_1, X_2, X_3 \rangle} = \frac{\langle \eta_1 | \eta_2 \rangle \langle \eta_3 | \eta_2 \rangle \langle \eta_2 \rangle \langle \xi_1 | \xi_2 \rangle \langle \xi_3 | \xi_2 \rangle \langle \xi_2 \rangle}{\langle \eta_2, X_1, X_2, X_3 \rangle} \\
&= \frac{\langle \eta_1 | \eta_2 \rangle \langle \xi_1 | \xi_2 \rangle \langle \eta_3 | \eta_2 \rangle \langle \xi_3 | \xi_2 \rangle \langle \eta_2 \rangle \langle \xi_2 \rangle}{\langle \eta_2, X_1, X_2, X_3 \rangle} \\
&= \frac{\langle \eta_1, X_1 | \eta_2, X_2 \rangle \langle \eta_3, X_3 | \eta_2, X_2 \rangle \langle \eta_2, X_2 \rangle}{\langle \eta_2, X_1, X_2, X_3 \rangle} \\
&= \frac{\langle \eta_1, X_1 | \eta_2, X_2 \rangle \langle \eta_3, X_3 | \eta_2, X_2 \rangle}{\langle X_1, X_3 | \eta_2, X_2 \rangle} = \ldots \\
&= \frac{\langle \eta_1, X_1 | \eta_2, X_2 \rangle \langle \eta_3, X_3 | \eta_2, X_2 \rangle}{\langle X_1 | \eta_2, X_2 \rangle \langle X_3 | \eta_2, X_2 \rangle} \\
&= \langle \eta_1 | \eta_2, X_1, X_2 \rangle \langle \eta_3 | \eta_2, X_2, X_3 \rangle.
\end{aligned}$$

The above is an illustration on how we prove (4.11) and (4.12) and we omit the lengthy proof of the general case. On the other hand, these expressions can be easily proved via the conditional independence property of a graphical model as in Lauritzen (1996).

A.3 Explicit expression for (4.12)

In the above subsection, we showed that:

$$\begin{aligned}
&\langle \boldsymbol{\eta}(t_k + 1 : t_{k+1} - 1) | \eta(t_k), \eta(t_{k+1}), \boldsymbol{\beta}, \mathbf{X}, \mathbf{Y} \rangle = \\
&\qquad \langle \boldsymbol{\eta}(t_k + 1 : t_{k+1} - 1) | \eta(t_k), \eta(t_{k+1}), \boldsymbol{\beta}, \mathbf{X}(t_k : t_{k+1}) \rangle.
\end{aligned}$$

We derive the explicit expression of the right-hand side above. First, it is easy to use the conditional independence property to find,

$$\begin{aligned}
&\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},\mathbf{X}(t_k:t_{k+1})\rangle \\
&\qquad =\frac{\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},X(t_k),X(t_{k+1})\rangle}{\langle \mathbf{X}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},X(t_k),X(t_{k+1})\rangle} \\
&\qquad =\frac{\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1})\rangle}{\langle \mathbf{X}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},X(t_k),X(t_{k+1})\rangle}.
\end{aligned}$$

Define two new variables,

$$\begin{aligned}
\boldsymbol{\eta}^c(t_k+1:t_{k+1}-1) =& \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}), \text{and} \\
\mathbf{X}^c(t_k+1:t_{k+1}-1) =& \mathbf{X}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},X(t_k),X(t_{k+1}),
\end{aligned}$$

such that

$$\begin{aligned}
&\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\eta(t_k),\eta(t_{k+1}),\boldsymbol{\beta},\mathbf{X}(t_k:t_{k+1})\rangle \\
&\qquad =\langle \boldsymbol{\eta}^c(t_k+1:t_{k+1}-1)|\mathbf{X}^c(t_k+1:t_{k+1}-1)\rangle. \qquad \text{(A.4)}
\end{aligned}$$

With the basic properties of the Brownian Bridge, it is easy to show that

$$\boldsymbol{\eta}^c(t_k+1:t_{k+1}-1) \sim \text{MVN}(\mathbf{f}_k, \sigma_H^2 \mathbf{R}_k), \qquad \text{(A.5)}$$

where

$$\begin{aligned}
f_k(t) =& f(t) + a_k'(t)(\eta(t_k) - f(t_k)) + a_k(t)(\eta(t_{k+1}) - f(t_{k+1})) \qquad \text{(A.6)} \\
=& a_k'(t)\eta(t_k) + a_k(t)\eta(t_{k+1}), \qquad \text{(A.7)} \\
a_k(t) =& \frac{t-t_k}{t_{k+1}-t_k}, \\
a_k'(t) =& 1 - a_k(t) = \frac{t_{k+1}-t}{t_{k+1}-t_k}, \\
R_k(s,t) =& \frac{(s-t_k)(t_{k+1}-t)}{t_{k+1}-t_k}, \qquad t_k < s \le t < t_{k+1},
\end{aligned}$$

and MVN$(\cdot,\cdot)$ denotes a multivariate normal distribution. The definition of $f(t) = A + (B-A)t/T$ is used to simplify (A.6) into (A.7) as follows,

$$\begin{aligned} a'_k(t)f(t_k) + a_k(t)f(t_{k+1}) =& \frac{t_{k+1}-t}{t_{k+1}-t_k}\left(A + (B-A)\frac{t_k}{T}\right) + \frac{t-t_k}{t_{k+1}-t_k}\left(A + (B-A)\frac{t_{k+1}}{T}\right) \\ =& A\frac{t_{k+1}-t_k}{t_{k+1}-t_k} + (B-A)\frac{(t_{k+1}-t)t_k + (t-t_k)t_{k+1}}{(t_{k+1}-t_k)T} \\ =& A + (B-A)\frac{(t_{k+1}-t_k)t}{(t_{k+1}-t_k)T} = A + (B-A)\frac{t}{T} \\ =& f(t). \end{aligned}$$

It leads to $f(t) - a'_k(t)f(t_k) - a_k(t)f(t_{k+1}) = 0$.

The definition of $\mathbf{X}^c$ and $\mathbf{X}$ in (4.5) implies that $\mathbf{X}^c$ is the sum of a deterministic term and two independent Gaussian processes:

$$\begin{aligned} \mathbf{X}^c(t_k+1:t_{k+1}-1) =& \mathbf{h}(t_k+1:t_{k+1}-1) + \boldsymbol{\eta}^c(t_k+1:t_{k+1}-1) \\ &+ \boldsymbol{\xi}^c(t_k+1:t_{k+1}-1), \end{aligned}$$

where $\mathbf{h}$ is defined in (4.5) and $\boldsymbol{\xi}^c$ is defined similarly as $\boldsymbol{\eta}^c$:

$$\boldsymbol{\xi}^c(t_k+1:t_{k+1}-1) = \boldsymbol{\xi}(t_k+1:t_{k+1}-1)|\xi(t_k),\xi(t_{k+1}) \sim \text{MVN}(\mathbf{g}_k, \sigma_D^2\mathbf{R}_k)$$

with $g_k(t) = a'_k(t)\xi(t_k) + a_k(t)\xi(t_{k+1})$.

In this way, the marginal distribution of $\mathbf{X}^c$ is

$$\mathbf{X}^c(t_k+1:t_{k+1}-1) \sim \text{MVN}(\mathbf{u}_k, (\sigma_H^2 + \sigma_D^2)\mathbf{R}_k), \tag{A.8}$$

where

$$\begin{aligned} u_k(t) &= h(t) + f_k(t) + g_k(t) \\ &= h(t) + a'_k(t)(\eta(t_k) + \xi(t_k)) + a_k(t)(\eta(t_{k+1}) + \xi(t_{k+1})) \\ &= h(t) + a'_k(t)(X(t_k) - h(t_k)) + a_k(t)(X(t_{k+1}) - h(t_{k+1})). \end{aligned}$$

Also, the covariance between $\boldsymbol{\eta}^c$ and $\mathbf{X}^c$ is

$$\mathrm{Cov}\left(\boldsymbol{\eta}^c, \mathbf{X}^c\right) = \mathrm{Cov}\left(\boldsymbol{\eta}^c, \boldsymbol{\eta}^c\right) = \sigma_H^2 \mathbf{R}_k. \tag{A.9}$$

From (A.5), (A.8), and (A.9), we can easily find that there is no need to invert any matrices when computing (A.4). As when the conditional covariance is calculated, the inverse of $\mathbf{R}_k$ will be canceled by $\mathbf{R}_k$, such that,

$$\mathrm{Cov}\left(\boldsymbol{\eta}^c, \mathbf{X}^c\right)\left(\mathrm{Cov}\left(\mathbf{X}^c\right)\right)^{-1} = \sigma_H^2 \mathbf{R}_k \left((\sigma_H^2 + \sigma_D^2)\mathbf{R}_k\right)^{-1} = \frac{\sigma_H^2}{\sigma_H^2 + \sigma_D^2} \mathbf{I}_{t_{k+1}-t_k-1}$$

$$\mathrm{Cov}\left(\boldsymbol{\eta}^c, \mathbf{X}^c\right)\left(\mathrm{Cov}\left(\mathbf{X}^c\right)\right)^{-1}\mathrm{Cov}\left(\mathbf{X}^c, \boldsymbol{\eta}^c\right) = \sigma_H^2\left(1 - \frac{\sigma_H^2}{\sigma_H^2 + \sigma_D^2}\right)\mathbf{R}_k = \frac{\sigma_H^2 \sigma_D^2}{\sigma_H^2 + \sigma_D^2}\mathbf{R}_k.$$

We only need to calculate $\rho = \sigma_H^2/\sigma_H^2 + \sigma_D^2$ for the conditional mean and covariance. In this way, the desired posterior in (A.4) is

$$\begin{aligned}
&\langle \boldsymbol{\eta}(t_k + 1 : t_{k+1} - 1) | \eta(t_k), \eta(t_{k+1}), \boldsymbol{\beta}, \mathbf{X}(t_k : t_{k+1}) \rangle \\
&\quad = \langle \boldsymbol{\eta}^c(t_k + 1 : t_{k+1} - 1) | \mathbf{X}^c(t_k + 1 : t_{k+1} - 1) \rangle \\
&\quad \sim \mathrm{MVN}\left(\mathbf{f}_k + \rho\left(\mathbf{X}(t_k + 1 : t_{k+1} - 1) - \mathbf{u}_k\right), \rho\sigma_D^2 \mathbf{R}_k\right).
\end{aligned} \tag{A.10}$$

A.4 Explicit expression of $[\phi, \beta, \eta_G | \mathbf{X}_G, \mathbf{Y}]$

Following the notation in Appendix A.1

$$\begin{aligned}
[\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\eta}_G | \mathbf{X}_G, \mathbf{Y}] &\propto [\boldsymbol{\phi}][\boldsymbol{\eta}_G | \boldsymbol{\phi}][\mathbf{Y} | \boldsymbol{\eta}_G][\mathbf{X}(t_{1:K}) | \boldsymbol{\beta}, \boldsymbol{\eta}(t_{1:K}), \boldsymbol{\phi}] \\
&\propto [\boldsymbol{\phi}] |\mathbf{R}_G|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\eta}_G - \mathbf{f}_G)^T \mathbf{R}_G^{-1} (\boldsymbol{\eta}_G - \mathbf{f}_G)\right\} \\
&\quad \times |\mathbf{D}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{Y} - \boldsymbol{\eta}_G)^T \mathbf{D}^{-1} (\mathbf{Y} - \boldsymbol{\eta}_G)\right\} \\
&\quad \times |\mathbf{C}_G|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{X}_G - \mathbf{Z}_G\boldsymbol{\beta} - \boldsymbol{\eta}_G)^T \mathbf{C}_G^{-1} (\mathbf{X}_G - \mathbf{Z}_G\boldsymbol{\beta} - \boldsymbol{\eta}_G)\right\}
\end{aligned} \tag{A.11}$$

where $\mathbf{X}_G = \mathbf{X}(t_{0:K})$ and $\mathbf{R}_G = \mathbf{R}(t_{0:K}, t_{0:K})$. Similar notation applies to $\boldsymbol{\eta}_G, \mathbf{f}_G, \mathbf{C}_G, \mathbf{Z}_G$. Also, we will introduce the following notation, which is similar to those in Appendix A.1, specifically the sub-vector or sub-matrix of those in Appendix A.1

with respect to the GPS observations:

- $\boldsymbol{\zeta}_G = (\boldsymbol{\beta}^T, \boldsymbol{\eta}_G^T)^T$, which is the joint vector of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}_G$ of length $Q+K-1$;
- $\mathbf{U}_G$ is a $K \times (Q+K-1)$ matrix with $\mathbf{U}_G(1:K, 1:Q) = \mathbf{Z}_G$, $\mathbf{U}_G(1:(K-1), (Q+1):(Q+K-1)) = I_{K-1}$ and $\mathbf{U}_G(K, (Q+1):(Q+K-1)) = 0$, which maps $\mathbf{U}_G\boldsymbol{\zeta}_G = \mathbf{h}_G + \boldsymbol{\eta}_G$;
- $\mathbf{V}_G$ is a $(K-1) \times (Q+K-1)$ matrix with $\mathbf{V}_G(1:(K-1), 1:Q) = \mathbf{0}$ and $\mathbf{V}_G(1:(K-1), (Q+1):(Q+K-1)) = \mathbf{I}_{K-1}$, such that $\mathbf{V}_G\boldsymbol{\zeta}_G = \boldsymbol{\eta}_G$;

Some algebra simplifies (A.11) to yield

$$
\begin{aligned}
&[\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\eta}_G | \mathbf{X}_G, \mathbf{Y}] \propto \\
&\quad \frac{1}{\sigma_H^2}\frac{1}{\sigma_D^2} \times |\mathbf{R}_G|^{-1/2} |\mathbf{D}|^{-1/2} |\mathbf{C}_G|^{-1/2} \times \\
&\quad \exp\left\{ -\frac{1}{2}\left[(\boldsymbol{\zeta}_G - \mathbf{M}_{G1}^{-1}\mathbf{M}_{G2})^T \mathbf{M}_{G1} (\boldsymbol{\zeta}_G - \mathbf{M}_{G1}^{-1}\mathbf{M}_{G2}) + M_{G3} - \mathbf{M}_{G2}^T \mathbf{M}_{G1}^{-1} \mathbf{M}_{G2} \right] \right\},
\end{aligned}
\tag{A.12}
$$

where

$$
\begin{aligned}
\mathbf{M}_{G1} =& \mathbf{V}_G^T \mathbf{R}_G^{-1} \mathbf{V}_G + \mathbf{V}_G^T \mathbf{D}^{-1} \mathbf{V}_G + \mathbf{U}_G^T \mathbf{C}^{-1} \mathbf{U}_G, \\
\mathbf{M}_{G2} =& \mathbf{V}_G^T \mathbf{R}_G^{-1} \mathbf{f}_G + \mathbf{V}_G^T \mathbf{D}^{-1} \mathbf{Y} + \mathbf{U}_G^T \mathbf{C}_G^{-1} \mathbf{X}_G, \text{and} \\
M_{G3} =& \mathbf{f}_G^T \mathbf{R}_G^{-1} \mathbf{f}_G + \mathbf{Y}^T \mathbf{D}^{-1} \mathbf{Y} + \mathbf{X}_G^T \mathbf{C}_G^{-1} \mathbf{X}_G.
\end{aligned}
$$

Following (A.12), it is easy to show that

$$
[\boldsymbol{\zeta}_G | \mathbf{X}_G, \mathbf{Y}, \boldsymbol{\phi}] \sim \text{ MVN}\left(\mathbf{M}_{G1}^{-1}\mathbf{M}_{G2}, \mathbf{M}_{G1}^{-1} \right). \tag{A.13}
$$

Integrating $\boldsymbol{\zeta}_G$ out, we get

$$
\begin{aligned}
[\boldsymbol{\phi} | \mathbf{X}_G, \mathbf{Y}] \propto & \frac{1}{\sigma_H^2}\frac{1}{\sigma_D^2} |\mathbf{R}_G|^{-1/2} |\mathbf{D}|^{-1/2} |\mathbf{C}_G|^{-1/2} |\mathbf{M}_{G1}|^{-1/2} \\
& \times \exp\left\{ -\frac{1}{2}\left[M_{G3} - \mathbf{M}_{G2}^T \mathbf{M}_{G1}^{-1} \mathbf{M}_{G2} \right] \right\}.
\end{aligned}
\tag{A.14}
$$

A.5 Marginal distribution of η at the non-GPS points

With (A.13), we can marginalize $\boldsymbol{\eta}_G, \boldsymbol{\beta}$ out in (A.10) to obtain $\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\mathbf{X},\mathbf{Y}\rangle$, which is later used in the numerical integration. Let $\boldsymbol{\zeta}_k = (\boldsymbol{\beta}^T, \eta(t_k), \eta(t_{k+1}))^T$ and its mean and covariance matrix obtained in (A.13) be denoted as $\tilde{\boldsymbol{\zeta}}_k, \tilde{\boldsymbol{\Sigma}}_k$ respectively.

The $\mathbf{u}_k$ of (A.8) can be written as as a linear transformation of $\boldsymbol{\zeta}_k$, such that $\mathbf{u}_k = \mathbf{B}_k\boldsymbol{\zeta}_k$, where $\mathbf{B}_k = \left[\rho\left(\mathbf{Z}_k - \mathbf{A}_k\mathbf{Z}_k^G\right), \mathbf{A}_k\right]$. $\{\mathbf{Z}_k\}$ are the rows of the design matrix, $\mathbf{Z}(t_k+1:t_{k+1}-1, 1:Q)$, which corresponds to this period of the non-GPS observation. $\mathbf{Z}_k^G = \mathbf{Z}(\mathbf{t}_{k,k+1}, 1:Q)$ corresponds to the rows of the design matrix of the two GPS observations. $\mathbf{A}_k$ is the matrix of the linear weights of $a_k'(t), a_k(t)$ as in Equation (A.5). Marginalizing $\boldsymbol{\zeta}_k$ out in (A.10) results in

$$\langle \boldsymbol{\eta}(t_k+1:t_{k+1}-1)|\mathbf{X},\mathbf{Y}\rangle \sim \text{MVN}\left(\mathbf{B}_k\tilde{\boldsymbol{\zeta}}_k, \mathbf{B}_k\tilde{\boldsymbol{\Sigma}}_k\mathbf{B}_k^T\right). \tag{A.15}$$

A.6 Integration over the variance parameters ϕ

In the BM literature (Liu et al., 2011), the integral in (4.15) was usually calculated by MCMC. However, the heavy computational burden of the MCMC technique has to be avoided in our application. Moreover, it may not be practical to store the MCMC samples of the high dimensional parameter $\boldsymbol{\eta}$. The first alternative is to avoid the integration via the empirical Bayesian approach as in Casella (1985):

$$[\boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{X},\mathbf{Y}] \approx [\boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{X},\mathbf{Y},\hat{\boldsymbol{\phi}}],$$

where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimate of $\boldsymbol{\phi}$ after $\boldsymbol{\eta}, \boldsymbol{\beta}$ are marginalized out,

$$\hat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} \log([\boldsymbol{\phi}|\mathbf{X}_G,\mathbf{Y}]). \tag{A.16}$$

The empirical Bayesian approach is computationally simple, especially when we can explicitly evaluate the marginal likelihood. However, it fails to reflect the uncertainty in $\boldsymbol{\phi}$ and thus it underestimates the uncertainty in the posterior of $\boldsymbol{\eta}$. To overcome this issue, we use a numerical integration method like that in INLA (Rue et al., 2009), which approximates the integral on a grid decided by the likelihood

surface.

Let $\mathbf{H}$ denote the 2×2 Hessian matrix of $\hat{\boldsymbol{\phi}} = (\hat{\sigma}_H^2, \hat{\sigma}_D^2)^T$ in (A.16) and $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$. With the eigenvalue decomposition $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T$, the space of $\boldsymbol{\phi}$ can be explored by $\boldsymbol{\phi}(\mathbf{z}) = \hat{\boldsymbol{\phi}} + \mathbf{A}\boldsymbol{\Lambda}^{1/2}\mathbf{z}$, where $\mathbf{z}$ is a 2×1 vector. To find the grid for numerical integration, we start from $\mathbf{z} = \mathbf{0}$ and search in the positive direction of z_1, that is, we increase $j \in \mathbb{N}^+$ and $\mathbf{z} = (j\delta_z, 0)$ as long as

$$\log([\boldsymbol{\phi}(\mathbf{0})|\mathbf{X}_G, \mathbf{Y}]) - \log([\boldsymbol{\phi}(\mathbf{z})|\mathbf{X}_G, \mathbf{Y}]) < \delta_\pi,$$

where δ_z is the step size and δ_π controls the magnitude of probability mass that will be included in the numerical integration. After searching on the positive side, we switch direction and search on the negative side of z_1. This procedure is repeated for both dimensions of $\mathbf{z}$. For our BM model above, if the search stops at J_1^+, J_2^+ steps in the positive directions of z_1 and z_2 respectively and J_1^-, J_2^- in their negative directions, a grid of size $(J_1^+ + J_1^- + 1) \times (J_2^+ + J_2^- + 1)$ is used in the numerical integration and the points on this grid are $\mathbf{z}_{j_1,j_2} = \delta_z(j_1, j_2)$, with $j_1 \in (-J_1^-, -J_1^- + 1, \ldots, -1, 0, 1, \ldots, J_1^+)$ and $j_2 \in (-J_2^-, -J_2^- + 1, \ldots, -1, 0, 1, \ldots, J_2^+)$. The integral in (4.15) is then approximated by

$$[\boldsymbol{\eta}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}] \approx \sum_{j_1=-J_1^-}^{J_1^+} \sum_{j_2=-J_2^-}^{J_2^+} w_{j_1,j_2} \times [\boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{\phi}(\mathbf{z}_{j_1,j_2}), \mathbf{X}, \mathbf{Y}], \tag{A.17}$$

where

$$w_{j_1,j_2} = \frac{[\boldsymbol{\phi}(\mathbf{z}(j_1, j_2))|\mathbf{X}_G, \mathbf{Y}]}{\sum_{j_1}\sum_{j_2}[\boldsymbol{\phi}(\mathbf{z}(j_1, j_2))|\mathbf{X}_G, \mathbf{Y}]}.$$

Note that (A.17) resembles Equation (5) of Rue et al. (2009). As the joint posterior distribution of $\boldsymbol{\eta}, \boldsymbol{\beta}$ conditional on $\boldsymbol{\phi}$ follows a multivariate Gaussian distribution, the posterior $[\boldsymbol{\eta}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}]$ can be approximated by a mixture of multivariate Gaussian densities, whose mean and variance can be easily calculated as follows.

The distribution of $\boldsymbol{\zeta}$ is multivariate normal conditional on $\boldsymbol{\phi}$ as in (A.13) and (A.15) and therefore the posterior density of $\boldsymbol{\zeta}$ can be approximated by a mixture of multivariate normal densities. Let $\tilde{\boldsymbol{\zeta}}^{(i)}$ and $\tilde{\boldsymbol{\Sigma}}^{(i)}$ be the posterior mean and covariance of $\boldsymbol{\zeta}$ conditional on the ith grid point of $\boldsymbol{\phi}(\mathbf{z}(j_1, j_2))$ (j_1, j_2 are collapsed

into a single index set and $L = (J_1^+ + J_1^- + 1) \times (J_2^+ + J_2^- + 1))$. We simplify (A.17) into

$$[\boldsymbol{\zeta}|\mathbf{X},\mathbf{Y}] \approx \sum_{i=1}^{L} w_i [\boldsymbol{\zeta}|\boldsymbol{\phi}^{(i)},\mathbf{X},\mathbf{Y}] = \sum_{i=1}^{L} w_i \Psi(\cdot;\tilde{\boldsymbol{\zeta}}^{(i)},\tilde{\boldsymbol{\Sigma}}^{(i)}),$$

where $\boldsymbol{\Psi}$ represents the probability density function of the multivariate normal distribution. For our application, we are only interested in finding the posterior mean and variance of $\boldsymbol{\zeta}$. As in (Frühwirth-Schnatter, 2006), the posterior mean equals

$$\tilde{\boldsymbol{\zeta}} = \sum_{i=1}^{L} w_i \tilde{\boldsymbol{\zeta}}^{(i)}, \tag{A.18}$$

and the posterior variance of the k-th element of $\boldsymbol{\zeta}$ is

$$\tilde{\sigma}_k^2 = \sum_{i=1}^{L} w_i \left[\tilde{\Sigma}^{(i)}(k,k) + \sum_{i=1}^{L} \left(\tilde{\zeta}^{(i)}(k) - \tilde{\zeta}(k) \right)^2 \right]. \tag{A.19}$$

Appendix B

Some Mathematical Details for CHBB

We provide the proofs and derivations for some equations in Section 5.2.2.

B.1 $|\mathbf{C}| = |\mathbf{C}_0||\mathbf{C}_{1,1}||\mathbf{C}_{1,2}|$

For a Brownian Bridge at time $\mathscr{T} = \{0,1,2,\ldots,T\}$ with variance parameter of 1, a decomposition of the joint distribution of $[\mathbf{y}_{1:T-1}|y_0,y_T]$ gives

$$\begin{aligned}[\mathbf{y}_{1:T-1}|y_0,y_T] =& [y_s|y_0,y_T][\mathbf{y}_{1:s-1}|y_0,y_s,y_T][\mathbf{y}_{s+1:T-1}|y_0,y_s,y_T] \\ =& [y_s|y_0,y_T][\mathbf{y}_{1:s-1}|y_0,y_s][\mathbf{y}_{s+1:T-1}|y_s,y_T]. \end{aligned} \tag{B.1}$$

We use the conditional independence property of BB to simplify the first line into the second line. Both sides of (B.1) are the normal densities

$$\begin{aligned}[\mathbf{y}_{1:T-1}|y_0,y_T] =& (2\pi)^{-(T-1)/2}|\mathbf{C}|^{-1/2}\exp\left\{(\mathbf{y}_{1:T-1}-\mathbf{f})^T\mathbf{C}^{-1}(\mathbf{y}_{1:T-1}-\mathbf{f})\right\} \\ [y_s|y_0,y_T] =& (2\pi)^{-1/2}|\mathbf{C}_0|^{-1/2}\exp\left\{(y_s-f_0)^T\mathbf{C}_0^{-1}(y_s-f_0\right\} \\ [\mathbf{y}_{1:s-1}|y_0,y_s] =& (2\pi)^{-(s-1)/2}|\mathbf{C}_{1,1}|^{-1/2}\exp\left\{(\mathbf{y}_{1:s-1}-\mathbf{f}_{1,1})^T\mathbf{C}_{1,1}^{-1}(\mathbf{y}_{1:s-1}-\mathbf{f}_{(1)})\right\} \\ [\mathbf{y}_{s+1:T-1}|y_s,y_T] =& (2\pi)^{-(T-s-1)/2}|\mathbf{C}_{1,2}|^{-1/2}\exp\left\{(\mathbf{y}_{s+1:T-1}-\mathbf{f}_{1,2})^T\mathbf{C}_{1,2}^{-1}(\mathbf{y}_{s+1:T-1}-\mathbf{f}_{1,2})\right\}, \end{aligned}$$

where $\mathbf{f}, f_0, \ldots$ are the mean functions of the Brownian Bridge as in (4.3) and all the other notations are the same as in Section 5.2.2.

Equation (B.1) holds for all $\mathbf{y}$. Let $\mathbf{y}_{0:T} = 0$, which leads to all $\mathbf{f} = 0, f_0 = 0, \mathbf{f}_{1,1} = 0, \mathbf{f}_{1,2} = 0$. The exponential part in the above normal densities are all 1. Plugging them back into (B.1), we have

$$\begin{aligned}(2\pi)^{-(T-1)/2}|\mathbf{C}|^{-1/2} =& (2\pi)^{-1/2}|\mathbf{C}_0|^{-1/2}(2\pi)^{-(s-1)/2}|\mathbf{C}_{1,1}|^{-1/2}(2\pi)^{-(T-s-1)/2}|\mathbf{C}_{1,2}|^{-1/2} \\ \Rightarrow |\mathbf{C}|^{-1/2} =& |\mathbf{C}_0|^{-1/2}|\mathbf{C}_{1,1}|^{-1/2}|\mathbf{C}_{1,2}|^{-1/2} \\ \Rightarrow |\mathbf{C}| =& |\mathbf{C}_0||\mathbf{C}_{1,1}||\mathbf{C}_{1,2}|\end{aligned}$$

B.2 Short BB are more likely to have a small variance estimate

As in (5.12), the MLE of the Brownian Bridge variance parameter is calculated as

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y}-\mathbf{f})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{f}),$$

where n denotes the number of free observations, i.e., $n = T-1$ for $BB(y_0, y_T, 0, T, \sigma_0^2)$. Notation σ^2 can denote any one in $\varsigma^2, \sigma_0^2, \sigma_{1,1}^2, \sigma_{1,2}^2$.

Without loss of generality, assume the true value $\sigma^2 = 1$. Because of the properties of multivariate normal distribution, $(\mathbf{y}-\mathbf{f})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{f})$ following χ_n^2 and therefore $\hat{\sigma}^2$ follows a Gamma distribution with shape $n/2$ and scale $2/n$. With this exact distribution, we can calculate the probability of $\hat{\sigma}^2 < z$, where z is a small number for different n. These probabilities are plotted in Figure B.1, which clearly shows that the probability of a small $\hat{\sigma}^2$ decreases with the increase of n.

B.3 Derivation of KL divergence for CHBB

The KL divergence for multivariate normal $P = \mathrm{N}(\mu_1, \mathbf{\Sigma}_1)$ and $Q = \mathrm{N}(\mu_2, \mathbf{\Sigma}_2)$ is

$$\mathrm{KL}(p||q) = \frac{1}{2}\left(tr\left(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0\right) + (\mu_1-\mu_2)\mathbf{\Sigma}_1^{-1}(\mu_1-\mu_2)^T - k - \log\left(\frac{|\mathbf{\Sigma}_0|}{|\mathbf{\Sigma}_1|}\right)\right)$$

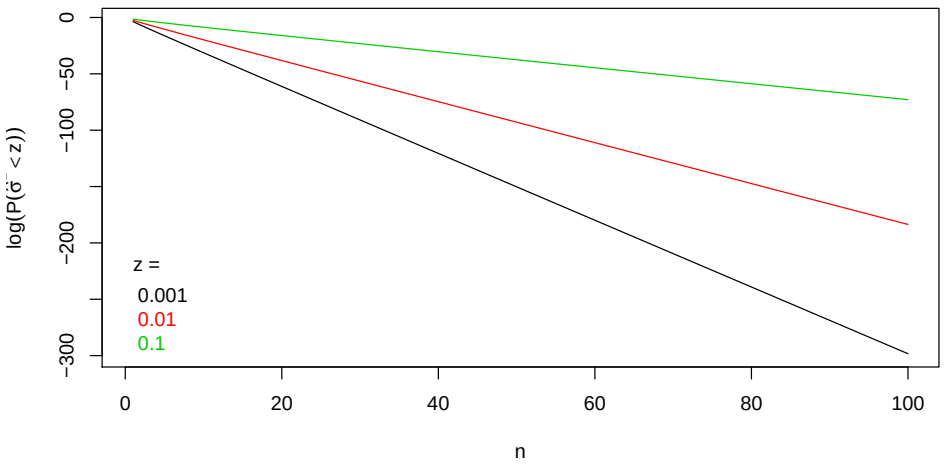


Figure B.1: $\log(P(\hat{\sigma}^2 < z))$ for different n and $z = 0.001, 0.01, 0.1$.

For $p = BB(y_0, y_T, 0, T, \varsigma^2)$ and $q = BB(y_0, y_T, 0, T, \sigma^2)$, the KL simplifies into

$$\mathrm{KL}(p||q) = (T-1)\frac{\varsigma^2}{\sigma^2} - (T-1) - (T-1)\log\left(\frac{\varsigma^2}{\sigma^2}\right). \tag{B.2}$$

Notice that the two BB with different variance parameters still have the same mean when they are defined on the same time points. Therefore $KL(p||q)$ does not involve the mean.

For the single knot CHBB in Section 5.2.2, we introduce the following notations for notational simplicity.

$$\begin{aligned}
p_0 &= [y_s | y_0, y_T, \varsigma^2] \\
q_0 &= [y_s | y_0, y_T, \sigma_0^2] \\
p_1 &= [\mathbf{y}_{1:s-1} | y_0, y_s, \varsigma^2] \\
q_1 &= [\mathbf{y}_{1:s-1} | y_0, y_s, \sigma_{1,1}^2] \\
p_2 &= [\mathbf{y}_{s+1:T-1} | y_s, y_T, \varsigma^2] \\
q_2 &= [\mathbf{y}_{s+1:T-1} | y_s, y_T, \sigma_{1,2}^2]
\end{aligned}$$

$$
\begin{aligned}
\mathrm{KL}(M_{(0)}||M_{(1)}) &= \int d\mathbf{y}_{1:T-1}[\mathbf{y}_{1:T-1}|y_0, y_T, \varsigma^2] \log\left(\frac{[\mathbf{y}_{1:T-1}|y_0, y_T, \varsigma^2]}{[\mathbf{y}_{1:T-1}|y_0, y_T, \sigma_0^2, \sigma_{1,1}^2, \sigma_{1,2}^2]}\right) \\
&= \int\int\int dy_s d\mathbf{y}_{1:s-1} d\mathbf{y}_{s+1:T-1} p_0 p_1 p_2 \log\left(\frac{p_0 p_1 p_2}{q_0 q_1 q_2}\right) \\
&= \int\int\int dy_s d\mathbf{y}_{1:s-1} d\mathbf{y}_{s+1:T-1} p_0 p_1 p_2 \log\left(\frac{p_0}{q_0}\right) \\
&\quad + \int\int\int dy_s d\mathbf{y}_{1:s-1} d\mathbf{y}_{s+1:T-1} p_0 p_1 p_2 \log\left(\frac{p_1}{q_1}\right) \\
&\quad + \int\int\int dy_s d\mathbf{y}_{1:s-1} d\mathbf{y}_{s+1:T-1} p_0 p_1 p_2 \log\left(\frac{p_2}{q_2}\right) \\
&= \int dy_s p_0 p_1 p_2 \log\left(\frac{p_0}{q_0}\right) + \int\int dy_s d\mathbf{y}_{1:s-1} p_0 p_1 \log\left(\frac{p_1}{q_1}\right) \\
&\quad + \int\int dy_s d\mathbf{y}_{1:s-1} p_0 p_2 \log\left(\frac{p_2}{q_2}\right) \\
&= \mathrm{KL}(p_0||q_0) + \int dy_s p_0 \mathrm{KL}(p_1||q_1) + \int dy_s p_0 \mathrm{KL}(p_2||q_2).
\end{aligned}
\tag{B.3}
$$

The last two terms in (B.3) only depends on the variance parameters but not the mean, which thus does not depend on y_s. This enables us to simplify (B.3) into,

$$
\mathrm{KL}(M_{(0)}||M_{(1)}) = \mathrm{KL}(p_0||q_0) + \mathrm{KL}(p_1||q_1) + \mathrm{KL}(p_2||q_2).
$$

Plugging (B.2) into the three terms,

$$
\mathrm{KL}(M_{(0)}||M_{(1)}) = g\left(\frac{\varsigma^2}{\sigma_{1,0}^2}\right) + (s-1)g\left(\frac{\varsigma^2}{\sigma_{1,1}^2}\right) + (T-s-1)g\left(\frac{\varsigma^2}{\sigma_{1,2}^2}\right),
$$

where

$$
g(x) = x - \log(x) - 1.
$$