State-space Models and Gaussian Processes in Paleoceanography and Paleoclimatology

A DISSERTATION PRESENTED
BY
TAEHEE LEE
TO
THE DIVISION OF APPLIED MATHEMATICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
APPLIED MATHEMATICS

BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND

MAY 2020

Abstract of "State-space Models and Gaussian Processes in Paleoceanography and Paleoclimatology", by Taehee Lee, Ph.D., Brown University, May 2020.

For the recent decades, statistical (machine) learning and Bayesian modelling have emerged as popular frameworks for various fields of area, regardless of academy or industry. The statistical learning theory has many applicable branches that include state-space models and the Gaussian process. State-space models are designed to deal with latent series under the Markov assumption. The hidden Markov model is applied for the case that each hidden state is defined on a finite set, whereas the Kalman filter, particle filter and Markov-chain Monte Carlo algorithms can treat the continuous cases. The Gaussian process gives a nonparametric prior on the functions. Its flexibility allows various applications including the regression, classification, state-space models and dimensionality reduction.

Paleoceanography and paleoclimatology are the fields of geoscience that discover the past ocean and climate events from the relevant preserved physical characteristics, proxies, as well as their ages. Interdisciplinary research encompassing physics, chemistry and biology has investigated the relationship between ocean and climate events and proxies, and more recently statistical learning is also making progresses. The scarce and agnostic characteristics of data need particular modelling that are more dependent on the data than certain parametric structures.

In this thesis, three applications of the state-space models and Gaussian process to the field of paleoceanography and paleoclimatology are presented. Dual proxy stack construction algorithm addresses the age assignments problem with the radiocarbon (^{14}C) and benthic $\delta^{18}O$ proxies by adopting a state-space model and the Gaussian process regression. The heteroscedastic Gaussian process regression constructs nonparametric calibration models of alkenone and TEX_{86} proxies for restoring past sea surface temperatures and a novel multimodal Gaussian process regression model is also considered. Two state-space models are applied for reconstructing past atmospheric CO_2 pressures with the boron isotope and benthic $\delta^{18}O$ proxies: one is a parametric state-space model based on the particle smoother and the other is a nonparametric Gaussian process state-space model. Our study certifies the effectiveness of giving correlation to the hidden states and considering nonparametric data-driven models in paleoceanography and paleoclimatology.

This dissertation by Taehee Lee is accepted in its present form by the Division of Applied Mathematics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ (Signature of Advisor)

Charles E. Lawrence, Advisor

Recommended to the Graduate Council

Date _____ (Signature of Reader)

Lorraine E. Lisiecki, Reader

Date _____ (Signature of Reader)

Matthew T. Harrison, Reader

Approved by the Graduate Council

Date _____ (Signature of Dean)

Andrew G. Campbell, Dean of the Graduate School

Curriculum Vitae

Taehee Lee received his Bachelor of Science in Mathematical Sciences with Financial Mathematics as a combined minor and Bachelor of Arts in Economics as a double major from Seoul National University in 2012. During his undergraduate studies he won the Presidential Science Scholarship from the Korea Student Aid Foundation (KOSAF) for 2007 to 2011. Once completing the alternative military service duty by 2014, he started the graduate studies in the Division of Applied Mathematics at Brown University. On September 17th, 2015, He was qualified to become a Ph.D. candidate by passing the preliminary oral exam. He worked as a teaching assistant (TA) for the courses 'Applied Ordinary Differential Equations' in Fall 2015 and 'Essential Statistics' in Spring 2016. Since September 2016, he has been working as a research assistant (RA) with Professor Charles E. Lawrence as his advisor. During his graduate studies he won the Kwanjeong Educational Foundation Doctoral Program Scholarship from the Kwanjeong Educational Foundation for 2014 to 2019. He also received two transitional master's degrees from Brown University in the interim: Master of Science in Applied Mathematics in 2017 and Master of Science in Computer Science in 2019. He was elected to SIGMA XI on May 16th, 2016 and is nominated for the Simon Ostrach Fellowship on May 24th, 2020. He successfully defended his Ph.D. thesis on April 16th, 2020.

Preface and Acknowledgments

When I was a high school student good at mathematics, I thought that the world can be explained objectively by a set of well-designed and abstract mathematical models. Such a belief led me to choose mathematics as the major in my undergraduate studies, and economics followed as a double major. Though I could graduate with an excellent academic record in 2012, I became more confused by the following philosophical doubt: can the inevitable subjectivistic viewpoint of the observer be always either ignored, removed, diluted or neutralized completely by a set of objective models? If so, is it okay to consider it to be subordinated to the objectivistic concept? For the next two years of my alternative military service as an assistant librarian in South Korea, I could fortunately read some classical books of the Austrian School and realized the importance of the subjectivistic aspects in theory and methodology. However, the subjectivistic aspects themselves do not explain but just rationalize the given results. How can we find a balance between the subjectivistic and objectivistic aspects in understanding the reality?

The Bayesian framework gives an ideal methodological answer to that question. Suppose that we have something unknown to discover, which is paired with the relevant known. The subjectivistic aspects that are independent from the known are formulated as a prior on the unknown while the objectivistic aspects appear as a set of models that defines the likelihoods of the known given the unknown. The Bayes' theorem derives the posterior of the unknown given the known from the prior and models. That is the reason why I decided to study the Bayesian statistical

learning in the Division of Applied Mathematics at Brown University.

Because most of the coursework in my undergraduate studies was not directly relevant to my current research areas, I could not avoid taking a lot of courses opened in the Division of Applied Mathematics or the Department of Computer Science after starting the Ph.D. program. To dig deep into the details of those courses, reading various textbooks and papers was essential. Weekly lab and personal meeting with my advisor comprise a core part of my studies. Statistical learning is indispensable from actual implementation, thus I needed to be familiar with some popular computer languages. Workshops, conferences, and seminars were good sources for the new ideas and recent research trends. Two transitional master's degrees, poster presentations, published and submitted papers and the award certify my sincere effort in the graduate studies.

My research areas cover a wide range of topics, from designing new statistical models to solving real problems. I picked some of the results from my research to organize this dissertation for proving that my graduate studies of the past six years are sufficient to grant me a Ph.D. degree in Applied Mathematics. To make my dissertation consistent, works in paleoceanography and paleoclimatology that mainly use state-space models and Gaussian processes were chosen. I organized them into three chapters according to their particular applications. Two additional chapters in Appendix were arranged to cover the general background of the models because I hope that my thesis will be also used as a manual for the beginners of the statistical learning theory.

During the graduate studies, I have received a lot of aid and support from many people. My thesis advisor, Prof. Charles Lawrence has contributed much to my studies. He has taught me the Bayesian statistical modelling, given inspiration for resolving difficult parts of the problems, connected me to interdisciplinary data-expert researchers that are relevant to my studies, discussed about new problems and modelling with me, introduced various workshops and conferences, supported expenses of my studies, and been the anchor whom I could rely on. I was also impressed by his gentle and kind manner to me. I indeed owe him the greatest deal of gratitude, next to my family.

I also appreciate the lab members. Prof. William Thompson has been providing good code

implementation. Seonmin Ahn and Taesoo Kim helped me a lot adapt myself to the new surroundings when I first arrived at Providence. Wilson McKerrow and August Guang were good seniors who suggested me proper papers and presented their works. Andrew Kaluzny and Daniel Kunin partially contributed to the works in section 4.3.1 and 2.3.1, respectively. Ashley Conard always actively shows and explains her work in computational biology: I regret not working a lot in research with her. I wish all of them good luck and success in their works.

Most of my research in paleoceanography and paleoclimatology could not be made without the aid of Prof. Lorraine Lisiecki and her graduate student, Devin Rand, at the University of California, Santa Barbara (UCSB). As researchers in paleoceanography, they have consistently collaborated with me by processing the raw data and analyzing my results qualitatively. Prof. Lorraine Lisiecki also participated in my dissertation defense as the second reader of the committee. Devin Rand has followed the rapid transition of my models and steadily given feedbacks to their results. Their efforts contributed a lot to the work in chapter 2. Dr. Geoffrey Gebbie in Woods Hole Oceanographic Institution also contributed to the work.

Prof. Matthew Harrison was willing to be participating in my dissertation defense as the third reader of the committee and be a reference for my job applications. He was also the instructor of the course 'Essential Statistics' for which I was a teaching assistant (TA) in Spring 2016. I appreciate his solicitude for me during the overall graduate studies of mine. I am also thankful for the efforts of Prof. Susan Gerbi and her lab members in the project of finding replication origins, though we eventually failed to obtain meaningful results. One of her students, John Urban, particularly spent much time for helping me understand their tools and methodology.

I am grateful to all the other faculty members, staffs, and graduate students in the Division of Applied Mathematics at Brown University. They have indeed contributed to making nice research environment. I also want to express my appreciation of the members of the Korean Graduate Student Association (KGSA) for their small help in my daily lives. Prof. Hyeong-in Choi of Seoul National University recommended me some noteworthy books of the Austrian School that eventually led me to change the majoring fields of studies from pure to applied mathematics.

For the first five years of my graduate studies, the Kwanjeong Educational Foundation had provided me with the Kwanjeong Educational Foundation Doctoral Program Scholarship. I am thankful to Chonghwan Lee, Honorary President of Samyoung Chemical Group who established the foundation. I also appreciate the public education system of Republic of Korea that focuses on not only granting sufficient opportunities to all young Korean students regardless of their families' wealth but also offering good special education for the gifted and talented ones.

Lastly, I need to express my gratitude to my family. My parents have always assisted my learning and studies. My maternal grandmother had sincerely aided me a lot with a keen interest before she passed away in 2012. My maternal grandfather financially supported my parents when I was very young and my father was just a graduate student. My maternal aunt has always been exceedingly supportive to me. I also appreciate all the other family members and relatives for their love.

Contents

1 Introduction **1**
- 1.1 Paleoceanography and Paleoclimatology . . . 2
- 1.2 Statistical Learning Theory . . . 6
- 1.3 Contents . . . 15
- 1.4 Dictionary . . . 17

2 Benthic δ^{18}O Stack Construction **21**
- 2.1 Related Work . . . 22
- 2.2 Research Motive and Data . . . 24
- 2.3 Core Alignment Algorithm . . . 26
- 2.4 Stack Construction Algorithm . . . 31
- 2.5 Results and Discussion . . . 34
- 2.6 Future Work . . . 42

3 Alkenone/TEX_{86} Calibration Models **44**
- 3.1 Related Work . . . 46
- 3.2 Research Motive and Data . . . 47
- 3.3 Models . . . 50
- 3.4 Results and Discussion . . . 57
- 3.5 Future Work . . . 66

4 Atmospheric CO_2 Reconstruction Using Boron Proxy **69**
- 4.1 Related Work . . . 70
- 4.2 Research Motive and Data . . . 71
- 4.3 Models . . . 72
- 4.4 Results and Discussion . . . 84
- 4.5 Future Work . . . 87

5 Conclusion **93**

Appendix A State-space Models **97**
- A.1 Hidden Markov Model . . . 100
- A.2 Kalman Filter and Smoother . . . 109
- A.3 Particle Filter and Smoother . . . 117
- A.4 Markov-chain Monte Carlo Algorithm . . . 122

Appendix B Gaussian Process Models **128**
B.1 Kernels 130
B.2 Gaussian Process Regression 138
B.3 Gaussian Process Classification 146
B.4 Gaussian Process State-space Models 151
B.5 Gaussian Process Latent Variable Models 155
B.6 Deep Gaussian Process 157
B.7 Variational Inference 159

References **185**

List of Tables and Illustrations

1.1 An example of how the indirect proxy is utilized for the age inference. Here, five sediment cores are aligned based on their fossil records. 3
1.2 A comparison of the frequentist and Bayesian perspectives. 8
1.3 A diagram of the parametric inference. The relationship between two variables X and Y are explained by the parametric model $f(\cdot|\theta)$ (left). The parameter θ is learned from the training data (X, Y) colored green (middle). The unknown Y is explained by the query X colored red and learned parameter θ (right). 11
1.4 Two nonparametric Gaussian process regressions on the same dataset but with different hyperparameters. 12
1.5 A diagram of the nonparametric inference. The relationship between two variables X and Y is not explained by any parametric model but is indirectly controlled by the hyperparameter θ (left). The hyperparameter θ is tuned from the training data (X, Y) colored green (middle). The unknown Y is explained by the query X colored red, tuned hyperparameter θ, and training data colored green (right). 13
1.6 Graphical representations of the generative model (left) and discriminative model (right). Here, θ is a parameter and X is a hidden variable that explains the observation Y. 15

2.1 Locations of cores GeoB7920-2, GeoB9508-5, GeoB9526-5, MD95-2042, MD99-2334, ODP658C, GIK13289-2 and SU90-08. 25
2.2 A graphical representation of our core alignment model. 26
2.3 Three steps for the stack construction algorithm. 31
2.4 A simple diagram of the SA-GPR. Each box is iterated until convergence. . . . 35
2.5 Core alignments to stacks. The upper panel (a) shows those to the DNA stack, while the middle panel (b) is our dual proxy local DNEA stack. Stars indicate medians of $\delta^{18}O$ data classified as inliers and dots represent outliers, after shifting them vertically by the core-specific shift parameters. The darker and brighter gray regions are the 1-sigma and 2-sigma of the stacks, respectively. The dot-dash line indicates their boundary. The lower panel (c) is a portion of the DNEA stack at 10-50 kiloyears. 36
2.6 A part of the DNEA stack constructed with the SE kernel. 37
2.7 Histograms of $\delta^{18}O$ to our dual proxy DNEA stack. Each $\delta^{18}O$ is standardized by the mean and standard deviation of the stack at the median of its sampled ages. 37

2.8 The comparison between inferred dual proxy ages and ^{14}C ages of records. Shaded regions indicate 95% confidence regions and the black dashed line is just a diagonal. 38
2.9 Dual proxy age inferences and alignments of GIK13289-2 to DNEA stack. In the upper figure, the darker and brighter areas show the 1-sigma and 2-sigma of the stacks, red points and blue dotted bars indicate medians and 95% confidence intervals of inlier age samples for $\delta^{18}O$ data, blue points and red dotted bars are the outliers. In the left below figures, cyan bars indicate 95% confidence intervals obtained independently from ^{14}C data, and blue and red areas show the 95% confidence bands of age inferences from ^{14}C only and both proxies, respectively. In the right below figures, medians (dashed lines) and 95% confidence intervals of relative ages to the medians of dual proxy ages are shown. 39
2.10 Dual proxy age inferences and alignments of SU90-08 to DNEA stack, similar to figure 2.9. This figure also shows how multiple observations at the same depths are dealt with in the alignment algorithm. 40
2.11 Histograms of the standardized accumulation rates. The blue histogram comes from the training ^{14}C dataset and the red one from the sampled age paths in constructing the DNEA stack. Thick lines indicate the inferred gamma distributions from the histograms, respectively. 42

3.1 $U_{37}^{K'}$ and TEX_{86} data. Each left panel shows the data locations while right plots data values over SSTs. 48
3.2 Some example results from our multimodal GPR. 58
3.3 Homoscedastic Gaussian process regression of $U_{37}^{K'}$ on SST. In each panel, red crosses are the outliers and shaded region indicates the 95% confidence band. . 60
3.4 Heteroscedastic Gaussian process regression of $U_{37}^{K'}$ on SST. 61
3.5 Multimodal Gaussian process regressions of $U_{37}^{K'}$ on SST. 62
3.6 Homoscedastic Gaussian process regression of TEX_{86} on SST. 63
3.7 Heteroscedastic Gaussian process regression of TEX_{86} on SST. 64
3.8 Multimodal Gaussian process regressions of TEX_{86} on SST. 65
3.9 Some preliminary results from the site-specific GPR calibration model. 67

4.1 A graphical representation of our particle smoother for the CO_2 reconstruction. 74
4.2 Ice volume and $\delta^{18}O$ pairs (red) and the chosen model with the 95% confidence band (green). 75
4.3 Atmospheric CO_2 and $\delta^{11}B$ pairs (red) and the chosen model with the 95% confidence band (green). 76
4.4 (Left) The BIC curve that selects 3 as the number of clusters and (Right) atmospheric CO_2 and ice volume index pairs (green) and the regressions of the clusters (in each color). 80
4.5 Initial and learned values of the parameters in the particle smoother model. The initial values were selected from [Garcia-Olivares & Herrero (2013)]. 85

4.6 Results from the particle smoother. (Top) Reconstructed atmospheric CO_2 with the experimental time series and published lower-upper bands. (Middle) Reconstructed extent of the Antarctic ice sheet indices. (Bottom) Reconstructed ice volume indices with the experimental time series from [Spratt & Lisiecki (2016)]. 86
4.7 Results from the Gaussian process state-space model with the SE kernel. 88
4.8 Results from the Gaussian process state-space model with the M25 kernel. . . 89
4.9 Results from the Gaussian process state-space model with the M15 kernel. . . 90
4.10 Results from the Gaussian process state-space model with the OU kernel. . . . 91

A.1 A graphical representation of the typical Bayesian inference example. Here the observations Y_n's are independent and identically distributed samples controlled by the parameter θ. 98
A.2 A graphical representation of the typical example of the state-space model. Here the observations Y_n's are conditionally independent given the hidden states X_n's. 100
A.3 Another type of the graphical representation of the hidden Markov model for the sequence alignment algorithm. Here, states S, E, I, D and M stand for start, end, insert, delete and match, respectively. 101
A.4 An example of HMM and the inference by the Viterbi and forward-backward algorithm. 106
A.5 An example of Kalman smoother where the 2-dimensional noisy observations are filtered. For each panel, black circles are the observations, the green curve is the true denoised observations (hidden states) and red curve is the inferred hidden states from the noisy observations. The left and right panels use the same observations except for 5% of the contamination in the right panels. 112
A.6 Examples of the particle smoother and mixture Kalman smoother on the same 2-dimensional dataset. Black circles are the noisy observations and green and red curves are the true and inferred denoised hidden states, respectively. 121

B.1 Functions sampled from the Gaussian process with different stationary kernels. 135
B.2 Functions sampled from the Gaussian process with different nonstationary kernels. 137
B.3 A graphical representation of GPR and GPC. Capital X and Y are the training inputs and outputs, respectively, and β indicates the latent state, that is either regression (GPR) or log-odd (GPC). Green x and red y are the query input and the associated output, respectively, and blue b stands for the associated latent variable. 138
B.4 Examples of the Gaussian process regression. Panel (a) and (c) are the homoscedastic cases whereas (b) and (d) are for the heteroscedastic data. 139
B.5 Examples of the GP classification. Each panel consists of four sub-panels. The first sub-panel show the artificial data that are drawn from the distribution depicted in the third sub-panel with 10% contamination. The second and fourth sub-panels show the posterior distributions of the labels with representing their levels with the brightness and the corresponding estimation, respectively. . . . 152

B.6 A graphical representation of GPST. Capital T and Y are the training indices (time) and outputs, respectively, and $X^{(1)}$ and $X^{(2)}$ indicate the latent states, that are hidden states to infer as those of the state-space models in chapter A. Green t and red y are the query index (time) and the associated output, respectively, and blue $x^{(1)}$ and $x^{(2)}$ stand for the associated latent variables. 153
B.7 A graphical representation of DGP. X and Y are the training inputs and outputs, respectively, and β and γ are the latent states (upper-left). Each column vector of β is assumed to follow a GP with X as the input (upper-right). Each column vector of γ is then assumed to follow another GP with β as the input (lower-left). Finally, Y is modelled given γ (lower-right). 158

1

Introduction

Since the dawn of the era of big data, statistical learning (machine learning) has been indispensable to any areas involved with large datasets thus played a crucial role in the innovation and advancement by taking advantage of the fast computer breakthroughs and Bayesian methodologies. As an example, computational biology has shown a remarkable progress in efficiently processing high-throughput data over the past few decades. Computer vision, speech recognition, computational linguistics, robotics, bioinformatics, and marketing are the other good examples.

Geoscience is not an exception. Various subfields including geophysics are actively adopting machine learning tools to analyze the relevant data, such as seismic waves, streamflow, and geomagnetic data. Paleoceanography and paleoclimatology to infer ancient events from proxies are

also depending much on statistical learning a lot in interpreting data. This dissertation aims at broadening such research with state-space models and Gaussian process applications.

1.1 Paleoceanography and Paleoclimatology

Geoscience is a branch of science that addresses the physical and chemical constitution of the Earth and its atmosphere. It consists of the following four studies: lithosphere, hydrosphere, atmosphere and biosphere. Paleoclimatology studies various types of environmental evidence to understand the past climate of Earth. Scientists utilize the relevant recorders to infer past conditions: such records are called proxies. Such studies have figured out that the Earth's climate system can shift between extremely different states abruptly and examined the causes of past climate changes to unravel how much the recent climate changes can be explained by natural causes and human influences. Paleoceanography is the same sort of studies but related more to the oceans.

1.1.1 Age Inference Problems

To reconstruct the past oceanic and climate events, one firstly needs to obtain their age information. The problem is that ages are also not measurable as well. In the field of geochronology, age inference is dependent on the proxies, just as the oceanic and climate events.

Age proxies are classified into two categories. Firstly, direct age proxies are allowing to date records for themselves. For example, ^{14}C determinations are assumed to follow the models based on radiocarbon ages, and such radiocarbon ages are translated into calendar ages by calibration curves [Reimer et al. (2013); Hogg et al. (2013)], thus one can access calendar ages via ^{14}C proxies. We call ages inferred by direct age proxies direct ages. Utilizing direct age proxies is dependent on their relationship with ages. If one wants to infer the ages of a sediment core, then the inference is also relevant to the relationship between ages and the associated layer depths.

Direct age proxies are often limited in quantity in each record and only available up to certain age ranges, thus we need complementary methods. Indirect age proxies are the global climate parameters (e.g., benthic δ^{18}O) that can be utilized as reference for aligning records which have

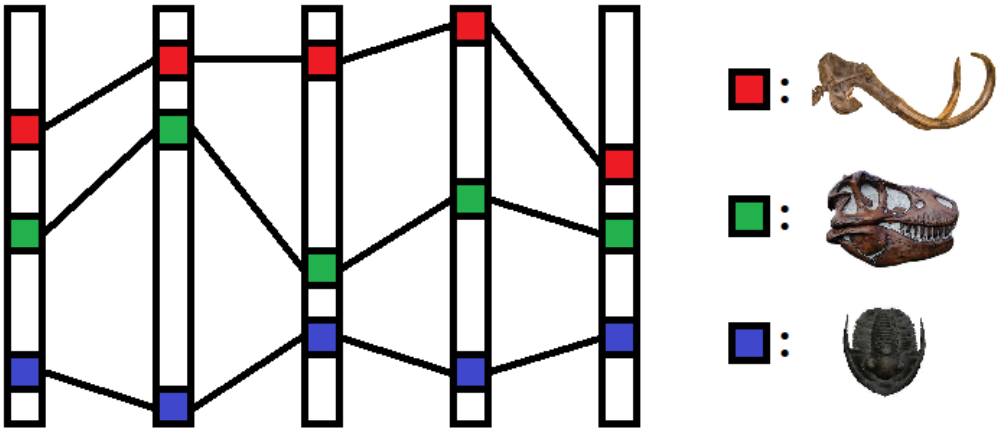


Figure 1.1: An example of how the indirect proxy is utilized for the age inference. Here, five sediment cores are aligned based on their fossil records.

direct proxies. Once records are aligned, direct ages at each record are shared with the others. Figure 1.1 shows an example. Utilizing indirect age proxies can be thus understood as the multiple sequence alignment problem. The dynamic time warping (DTW) [Munich & Perona (1999)] is not sufficient to address this problem because it does not exclude the extreme alignments based on the layer depths and the level of sedimentation rates. Also, pairwise alignment cannot be extended straightforwardly to the multiple sequence alignment because it does not guarantee consistency of the alignments and the time complexity is quadratic of the number of sequences.

Both direct and indirect age proxies can be utilized in the Bayesian framework. The likelihood is consisting of the proxy models given ages, while the prior on ages reflects the prior knowledge from the relationship between ages and secondary information, such as layer depths and sedimentation rates. From the likelihood and prior, one can derive the posterior distribution of ages by the Bayes' theorem. One can simultaneously use the multiple kinds of age proxies straightforwardly from their proxy models if they are conditionally independent given ages. We will further discuss about this topic in chapter 2.

1.1.2 Reconstruction Problems

Paleoceanography and paleoclimatology investigate the ocean events and past climate, respectively. We have no way of measuring them directly. Instead, the inference can be done by the

proxy methods that utilize the causality or correlation between some particular preserved physical characteristics and those events. There are various types of proxies.

Mountain glaciers and the polar ice caps and sheets are good sources for the data. The trapped air is a source for direct measurement of the composition of air. Changes in the layering thickness are connected to those in precipitation or temperature. $\delta^{18}O$ in ice layers represent changes in average sea surface temperature. Sediments contain remnants of preserved vegetation, animals, plankton, or pollen that characterize certain climate zones. Biomarker molecules, chemical signature and isotopic ratios carry further information. Coral rings respond to water temperature, freshwater influx, pH changes and wave action.

Climate and ocean events are highly correlated one another. Proxies could be correlated with multiple events. Statistical inference on them without assuming the underlying theory would be unreliable, especially if only a small amount of data is given. Instead, the theory gives a set of equations that represent the relationships: for example, [Dickson (2010)] characterizes the ocean carbon system with several equilibrium equations. Deriving the explicit calibration model for the climate events to infer given proxies is done by solving the system of those equations. If it is hard to derive such an explicit model, sample-based methods by Markov-chain Monte Carlo algorithm can be considered to be an alternative.

If such a theory is unavailable, then the data-driven methods would be the second-best choice. Here, to avoid overfitting that comes from the curse of dimensionality, either simple parametric models or nonparametric methods applied on only one or two kinds of proxies and events are considered. For example, [Epstein et al. (1953)] formulates the relationship between ice volumes and $\delta^{18}O$ as a quadratic extrapolation. Though this approach would seem to be limited and ignorant of the other events and proxies at first glance, it gives another way of utilizing multiple proxies by adopting those models as observation distributions in the framework of Bayesian inference.

In either case, the proxy methods depend on the present-day environmental conditions to make statements about past conditions, which is called "uniformitarianism principle" [Scott (1963)]. If the relationships can be explained theoretically based on the fields of physics and/or

chemistry, then it is not too much a stretch to assume it; if the proxies are involved with biological processes, however, then nonstationary issues that stem from the evolution can be problematic. Each proxy has its own span and resolution that characterize its availability: for example, the ^{14}C proxy is limited up to 50 kiloyears and tends to have a low resolution in a record.

This dissertation also contributes to the reconstruction of sea surface temperatures (chapter 3) with alkenone and TEX_{86} proxies and atmospheric CO_2 concentrations (chapter 4) with benthic $\delta^{18}O$ and boron proxies.

1.1.3 Characteristics of the Data

For the last decades, machine learning (or statistical learning) has become very popular in various research areas, as long as analyzing data gives more opportunity. Bayesian analysis and, more recently, deep neural networks [Goodfellow et al. (2016)] have led to innovations. This dissertation also aims at such a purpose in paleoceanography and paleoclimatology. Before applying a model, it is useful to discuss about the characteristics of our data.

Data in paleoceanography and paleoclimatology are consisting of various proxies that have been imprinted for the geological time, from a few thousands to hundreds of millions of years. It means that generating such data is usually infeasible in laboratories: even if one somehow knows how to accelerate the processes, they are also affected by the past environmental conditions (such as salinity, temperatures, acidity, ocean circulations, etc.) that are not known a priori. Thus, data are limited in the observations from the existing records.

It is also difficult to gather such data as much as we want. Firstly, some records are expensive to process. For example, ocean sediment cores are extracted from the thousands of meter deep oceans. Secondly, processing proxies in a record could be either expensive or mostly infeasible due to the tiny amounts thus only a limited number of observations are made in practice. Thirdly, records of high quality are rare: contamination by non-climatic influences could be an issue.

For such reasons, data in paleoceanography and paleoclimatology are often limited in quantity. Also, some proxies have not been fully investigated yet so there is no accurate model to describe

the relationship between them and the associated past climate events. These limitations make the inference more loyal to the data and bring more cautious analysis to the outliers: if data are many, discarding any suspicious data in the preprocessing would not lose information much. However, if data are scarce and likely to include outliers, one needs to classify them more carefully for minimizing information loss.

1.2 Statistical Learning Theory

Statistical learning is a framework for the inference and prediction drawing from statistics and functional analysis [Hastie et al. (2009)]. Nowadays the term 'machine learning' is used more widely: it usually aims at accurate predictions so all about the results, while statistical learning tends to focus more on the relationship between variables. The blurred boundary between those two fields comes from the modern trend that machine learning algorithms are often highly depending on the statistical theory and statistical learning is not only limited to understanding the data but also prediction based on it.

However, they are not identical. Suppose that we have a huge amount of high-dimensional data to be exploited for the prediction. A typical approach of machine learning is to search the most predictive and/or robust model to the data, usually at the cost of interpretability. Some data-driven methods such as deep neural network models sometimes outright ignore the underlying data-specific theories that have been studied by experts. This approach could be good if we do not know much about the data and have urgent demands in engineering, while the deviation from the existing theory might make the experts hesitate to take it instead of the original models that could be suboptimal but are highly interpretable in theory.

In contrast, statistical learning is first pursuing the understanding of data probabilistically: it focuses on determining the causality and correlation between variables. Once a probabilistic model on the given data is settled, it is then utilized for deriving the prediction. Though statistical learning can be suboptimal, especially if the model for the data is misspecified, it facilitates the use of previously reported theory germane to the data. Thus, this approach might be better if the

reliability of models is determined by how much they reflect data-specific characteristics.

Because this dissertation research has the purpose in giving more efficient and effective models on the scientific data in paleoceanography and paleoclimatology, the methodology is inevitably more faithful to the theory. For example, it could be fancy to align two sediment records over ages by utilizing the proxy pattern only, but the model will never attract the experts if it does not count on the sedimentation rate models: extreme alignments based on the depths lead to concluding extreme accumulation in past that could be extremely unrealistic. We will discuss more about the methodology in the following subsections.

1.2.1 Frequentist vs. Bayesian Inference

Statistics is the academic field involved with the collection, organization, analysis, interpretation, and presentation of data. Furthermore, it deals with the uncertainty that stems from the data. Though the modern statistics is intensively written in the language of mathematics, regarding statistics as a branch of mathematics is deceptive. While mathematics is building a theory with definitions and theorems that can be applied to explain the phenomena, statistics aims at constructing processes for comparing and measuring hypotheses and models to deal with uncertainty. While mathematicians design an ideal setting and wish that it is close enough to explain the reality, statisticians are engrossed in creating models that can encompass its uncertain aspects.

Debates on the frequentist vs. Bayesian reasoning have a long history. Suppose that we have a model that addresses the given data. Frequentists consider that the model is certain but unknown and the associated data can be generated given the model. Thus, they consider that data are random whereas the model is not. Bayesians have a different point of view: they consider that the given data are themselves certain so not random. Instead, they consider models to be random (prior). Data are just explained given models (likelihood). By applying Bayes' theorem [Stuart & Ord (1994)], they measure the models given data (posterior). To be more abstract, the frequentists regard an event's probability as the limit of the relative frequency in multiple trials thus objective while the Bayesian consider it to be the degree of belief so subjective. Figure 1.2

	Frequentists	Bayesians
parameters	unknown constants	random variables
data	randomly generated from models	not random but given
probability	limit of the relative frequency	reasonable expectation or belief
source	observations	observations and prior belief
decision	hypothetical test	model selection via posteriors
goal	explaining the data uncertainty	updating the belief from data

Figure 1.2: A comparison of the frequentist and Bayesian perspectives.

gives a table to compare two reasonings. Note that the modern frequentists use Bayes estimators frequently because they have good frequentist properties: see [Wasserman (2004)]. Here we just discuss the underlying perceptions and philosophies of the frequentists and Bayesians.

One obstacle from the frequentist philosophy in the field of statistical learning is that the inference and prediction become more unrealistic and overfitting to the data as the number of parameters increases. Because they consider parameters to be unknown but certain values, parameter estimation implicitly assumes that the model is true. This approach could be effective if the model has been physically justified but what we do not know are just its parameters. However, it becomes less robust if the model itself is misspecified. To deal with it, frequentists often distinguish the term 'parameter' and 'hidden variable', where the latter is a random variable that hide in the model as a latent state. Bayesians, however, can treat parameters as random variables so are conceptually free to integrate them out without introducing additional concepts: model uncertainty can be addressed consistently by doing so.

Bayesians have been continuously criticized for their unscientific aspects: if probability is essentially *subjective*, then how can we achieve consensus on the inference and prediction given data? More technically, all results are depending on the choice of priors. If one set a strong prior, then the data will be dominated by our prior belief; if a weak or flat prior is given, then the model is just bothersome without any particular advantages [Gelman (2008)]. Without depending on the conjugate priors, the posteriors are often not possible to be expressed in closed forms. Moreover, posteriors as the conjugate distributions can usually be replicated by the frequentist approach with pseudo-counts on the data, thus not that exclusive to Bayesian, just as it circumvents the

certainty of parameters by introducing hidden variables.

Despite such shortcomings and criticism of Bayesian statistics, it is often both meaningful and practical in the statistical *learning*. Firstly, it is quite common that data are given a lot whereas their characteristics are mostly unknown in practical problems. If so, the models that are applied to the data are either misspecified or involved with too many parameters, thus the overfitting rises as an issue: this can be avoided by assigning particular values to the parameters, and Bayesian inference that justifies the marginalization of parameters is thus practical. In addition, here the models themselves are assumed not to be justified but somehow selected by the solver so "subjective beliefs" in the Bayesian literature would not more be conceptually problematic. Also, data-driven approach is for deriving models from the given data, thus it is more natural to assume that data are not random (as what frequentists say) but given (as Bayesians say) in that case.

Secondly, suppose that we somehow know the characteristics of data, but not sure, then having some candidates instead of a single model is desirable. Frequentist statistics deal with such a problem by the (statistical) hypothesis test. It starts with selecting the most probable and sensible model as the null hypothesis and defines the others to be the alternative hypothesis. Then it either accepts or rejects the null hypothesis based on the relevant test statistic obtained under the null hypothesis and a significant level (or p-value). This approach is quite annoying for the multiple model selection and implies that the tester already has a certain belief on the models: why is one supposed to be null and the other alternative? Bayesian model selection gives a succinct and explicit solution to this problem: it just computes the posteriors of models given their prior beliefs. Moreover, marginalizing models out to get the posterior predictive model is another option for reducing uncertainty.

Thirdly, in some real problems, it is probable that partial information of the parameters and variables to learn is available independent of the data. Bayesian framework gives clear means to deal with this: priors on the parameters and variables are interpreting such information. If so, the priors are not representing the *subjective* beliefs on the values to estimate anymore: they are now *objective* in some sense. Furthermore, Bayesian framework is more intuitive in this case:

researchers are always going forward from the prior knowledge of the objectives to study and supplementing the knowledge with the new findings from the research. This is equivalent to calculating the posterior of the unknown from the prior and likelihood.

Lastly, Bayesian approach could be more robust and stable in practice, especially if the data are scarce. In discrete problems, (a strict form of) frequentist approach might assign zero probabilities to some states: if no observation at a state is made, then it is estimated to never happens. If such possibilities are open, some useful and powerful results in the probability theory cannot be applied. For example, the Hammersley-Clifford theorem [Clifford (1990)] works under the assumption that probability distributions are strictly positive. Giving a nonzero prior can resolve such singularity at once. Also, Bayesian framework prevents the estimation from being sensitive to the initial conditions that might lead estimations to local optima, especially if the algorithm consists of iterations such as Gibbs sampling [Geman & Geman (1984)]. Here, priors are working as "anchors" so that the estimate or sampled parameters or hidden variables are not deviated much from the "probable" ranges.

For the above reasons, I have been working on my research topics as a Bayesian statistician, and this dissertation is also written in the Bayesian language. As discussed in section 1.1.3, data in paleoceanography and paleoclimatology are often lacking an underlying physical dynamics model compared to the other fields that statistical/machine learning are applied to. It is impossible or expensive to generate or gather such data. Thus, Bayesian philosophy that assumes data to be given is practical for this problem.

1.2.2 Parametric vs. Nonparametric Model

A statistical model is a set of distributions, densities, or regression functions [Wasserman (2004)]. Statistical models can be classified into the following two categories: parametric and nonparametric models. A parametric model can be parametrized by a finite number of parameters that take values in a certain parametric space, which is given independently from the data. A nonparametric model assumes that the data cannot be explained in terms of such a finite set of parameters

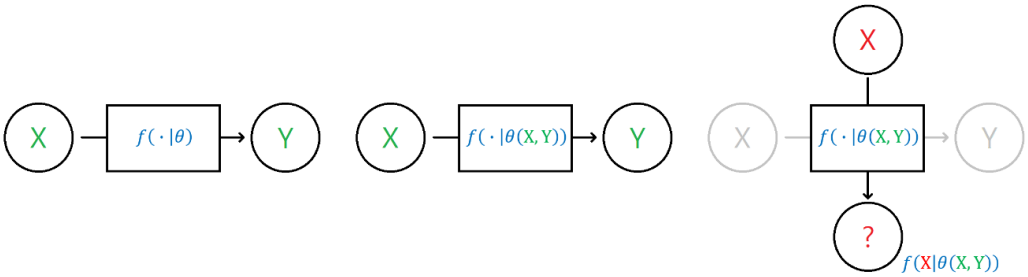


Figure 1.3: A diagram of the parametric inference. The relationship between two variables X and Y are explained by the parametric model $f(\cdot|\theta)$ (left). The parameter θ is learned from the training data (X, Y) colored green (middle). The unknown Y is explained by the query X colored red and learned parameter θ (right).

only. More intuitively, parametric models are described by the functions that are defined by the parameters only, where nonparametric ones are those by not only parameters (often called hyperparameters) but also data themselves to depend on the structural assumptions as few as possible [Wasserman (2006)].

Statistical learning with parametric models can be understood as inferring parameters from the data. Frequentists have some criteria for this procedure, such as the maximum likelihood estimator and (expected) loss, based on the training and validating data. Bayesians consider parameters to be random, so have other criteria for inferring parameters, such as maximum a posteriori probability estimator and Bayesian expected loss. Once a set of parameters are learned, the model is fixed because it is dependent on the parameters only. One more thing that Bayesians can do is to marginalize out those parameters over their posterior distributions to obtain a posterior predictive model: here the model is controlled by the prior hyperparameters. Though fast inference is possible once the parameters are learned, the model complexity is bounded so less flexible in this approach: it is not problematic if the model is investigated to describe the data exactly thus parameters are themselves meaningful, but could be problematic if the model is misspecified. If so, either overfitting or suboptimality could be an issue to deal with. Polynomial regression, logistic regression, finite mixture models, hidden Markov models, K-means, neural networks are examples of the parametric model. Figure 1.3 shows a diagram of the procedure for the parametric inference.

Though the nonparametric models are more data-dependent and free from the structural

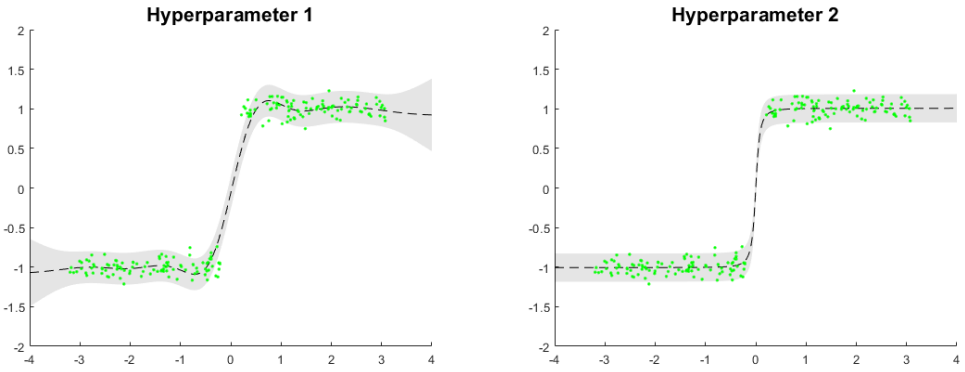


Figure 1.4: Two nonparametric Gaussian process regressions on the same dataset but with different hyperparameters.

models, they are controlled by hyperparameters: see figure 1.4. Thus, nonparametric learning is about tuning hyperparameters, such as bandwidths and number of clusters. One major difference from the parametric learning is that such hyperparameters can often not be estimated by maximizing likelihoods or minimizing losses: such attempts lead the model to be ignorantly loyal to the training data. Instead, they are "tuned" by optimizing objective functions that are regulated for avoiding the overfitting to the training data. Because the tuned model is still a function of data, inference could be done slowly if the time complexity is super-linear and data size is large. A nonparametric model is structural-free so more flexible than the parametric ones, thus useful if any parametric model is hypothetical. Moreover, it is controlled by a small number of hyperparameters thus can avoid overfitting from the lack of data compared to the number of parameters in the parametric models. However, nonparametric models could be more susceptible to suffer from the curse of dimensionality [Geenens (2011)] that means the number of training data to guarantee reliable inference increases tremendously as the dimension of predictors grows. Dirichlet process, Gaussian process, kernel density estimation, K-nearest neighbors and support vector machine with a positive definite kernel are examples of the nonparametric model. Figure 1.5 shows a diagram of the procedure for the nonparametric inference.

This dissertation mainly depends on the following two models: state-space models and the Gaussian process. The formers are mostly parametric whereas the latter is nonparametric. If reliable or justifiable models are given by the data-experts, then adopting parametric models is more

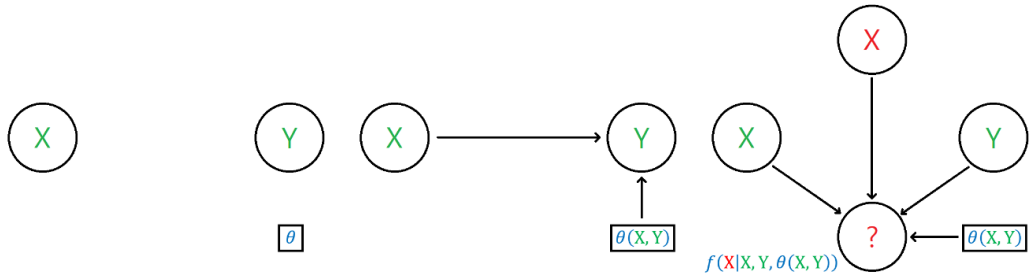


Figure 1.5: A diagram of the nonparametric inference. The relationship between two variables X and Y is not explained by any parametric model but is indirectly controlled by the hyperparameter θ (left). The hyperparameter θ is tuned from the training data (X, Y) colored green (middle). The unknown Y is explained by the query X colored red, tuned hyperparameter θ, and training data colored green (right).

advantageous in the data analysis and learning procedure. Otherwise, data-driven agnostic models could be considered. If the amount of available training data is large, then models based on the deep neural network have been reported to work well in practice. However, due to their excessive number of parameters, it would lead to overfitting to apply such models to the moderate number of data. In such cases, nonparametric models that are depending directly on the training data can be chosen for the data-driven purpose. Note that there is no panacea in statistical learning: the choice of models should be made based on the characteristics of data.

1.2.3 Generative vs. Discriminative Model

Statistical learning is a statistical attempt to find the relationship between variables. There are two types of variables: independent and dependent variables that can be understood as the observable inputs (predictors) and target outcomes resulting from varying inputs, respectively. Practically, what the statistical learning results in are the distribution function of the dependent variable (output) given independent one (input) from the training data. From now on the terms *observations* and *targets* will be used for expressing independent and dependent variables, respectively. Note that this classification is a bit hypothetical and goal-oriented: variables needs not to be the result of other variables.

A generative model starts with investigating the variables without such classification: in mathematical words, it aims at constructing the joint distribution of observations and targets that can

be utilized for deriving the conditional distribution of targets given observations by the Bayes rule. The definition is equivalent to the construction of conditional distribution of observations given targets if a prior on the targets is given in Bayesian terminology: "a target *generates* the associated observation". This approach is appropriate if 1) the structure of the relationship between the observations and targets are given a priori by experts so only the parameters are needed to be learned or 2) the conditional distribution of observations given targets is easy to deal with – for example, a hidden Markov model assumes that observations are conditionally independent given targets (hidden states). Generative models also give a way of generating pseudo-observations artificially. Mixture models, hidden Markov models, Bayesian networks, latent Dirichlet allocation, Boltzmann machines, variational autoencoders, and energy based-models are the examples of generative models.

A discriminative model is more pragmatic: it directly approximates the conditional distribution (or estimates, if the model is deterministic) of targets given observations. Such distributions are usually hypothetical and independent from the data-specific characteristics because observations are not actually generating the targets. This approach is usually effective if the training data are plentiful and targets are defined in a small finite set or continuous but low-dimensional. However, it either becomes intractable or suffers the curse of dimensionality especially if the targets are defined in a high-dimensional space and expected to be highly correlated one another. K-nearest neighbors, logistic regression, support vector machines, maximum-entropy Markov models and neural networks are the typical examples. Figure 1.6 gives the graphical representations of generative and discriminative models.

In paleoceanography and paleoclimatology, proxies are the raw observations and past climate events or ages are the targets to infer. Here it is reasonable to perceive that past climate events have been imprinting the proxies. Also, it is quite common that multidisciplinary research gives a meaningful prior on the past climate events. Some events, such as ages, are highly correlated one another. Thus, inferences in this dissertation are done in the framework of generative models.

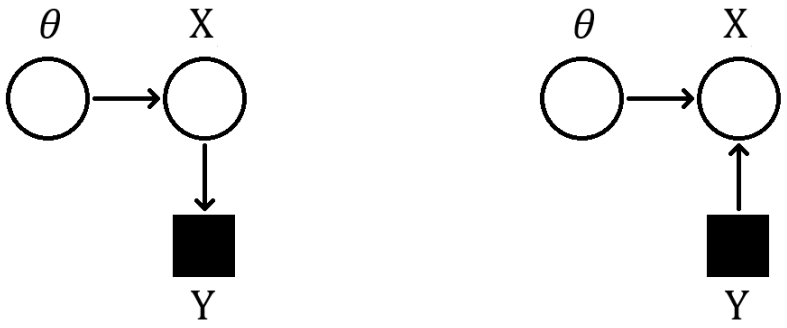


Figure 1.6: Graphical representations of the generative model (left) and discriminative model (right). Here, θ is a parameter and X is a hidden variable that explains the observation Y.

1.3 Contents

As discussed so far, chapter 1 covers the basic concepts and knowledge to give a departure point of the dissertation. It mainly consists of two parts. One is related to the application field, paleoceanography and paleoclimatology, about their issues to be addressed and characteristics distinctive from the other scientific fields. The other briefly introduces the statistical learning theory with some discussions on the methodologies that fit for the application area. Notation and symbols that continuously appear in the rest of the dissertation are covered in the latter part of this chapter.

Chapters 2, 3 and 4 present the applications to paleoceanography and paleoclimatology. Chapter 2 covers how an indirect age proxy, benthic δ^{18}O, is utilized together with a direct age proxy, ^{14}C, for inferring ages by the multiple sequence alignment algorithm. After briefly introducing the related work, it firstly shows the multiple core alignment algorithm given a reference (or motif/profile), based on the hybrid of particle smoother and Markov-chain Monte Carlo algorithm. Then it presents how to construct a stack that is used as the reference of the multiple core alignment algorithm given core alignments, based on the Gaussian process regression. How these two algorithms form a complete core alignment algorithm and related future work will be covered lastly.

Alkenone and TEX_{86} are proxies for reconstructing sea surface temperatures. Chapter 3 is about the calibration model construction for utilizing alkenone and TEX_{86} based on the heteroscedastic Gaussian process regression. Firstly, an extended version of the heteroscedastic Gaussian process regression from my published work is covered briefly. Then the detailed formulation of novel multimodal Gaussian process regression method is discussed. Lastly, we compare and analyze the constructed calibration models according to the introduced methodologies.

Some climate events are correlated so can be inferred simultaneously from the associated proxies. For example, atmospheric CO_2 pressure and ice volumes are correlated thus can be inferred together from the boron isotopes and benthic $\delta^{18}O$. Chapter 4 suggests two different models for this work. The first model uses the particle smoother to design the model based on the dynamic system, while the second one is derived by the Gaussian process state-space model. The comparison between the models and relevant future work will be discussed at the end of the chapter.

Finally, the dissertation ends with chapter 5 that covers the conclusion. In Appendix, two additional chapters cover the theoretical background used throughout this dissertation and contain detailed computations that are omitted in the main body to emphasize on our novel works. To be specific, several topics about state-space models are presented in Appendix A. It starts with discrete models such as the hidden Markov model and then deals with continuous but exact models such as the Kalman filter and smoother. Because such exact models are quite limited in some specific cases, approximation methods based on the sampling such as particle filter and smoother and Markov-chain Monte Carlo algorithm follow.

Appendix B discusses Gaussian processes and their applications in statistical learning. After defining some relevant concepts, it classifies kernel functions that determine the performance. Then the applications of regression, classification, state-space models, and dimensionality reduction follow. Lastly, the deep Gaussian process and variational inference are covered to handle the drawbacks of Gaussian processes that stem from the suboptimality and time complexity.

1.4 Dictionary

This section lists the notation, symbols and acronyms that are used in the rest of my dissertation. Acronyms are redefined at every chapter.

1.4.1 Notation

A Gaussian process needs the mean and kernel covariance functions with data input X that could be multidimensional, i.e., it has multiple features. We assume that each row of X represents the input datum. For example, if X is given as a $N \times D$ matrix, then it has N data and each datum is D-dimensional. To reflect its consistent extensibility, this dissertation abuses some notations. Let μ be the mean function. Then, for any $N \times D$ matrix X, $\mu(X)$ is the $N \times 1$ vector where its nth entry is $\mu(X_n)$. Similarly, for a kernel covariance function $\mathbb{K}$ and matrices $X^{(1)}$ and $X^{(2)}$ that have sizes $N_1 \times D$ and $N_2 \times D$, respectively, $\mathbb{K}\left(X^{(1)}, X^{(2)}\right)$ represents the kernel covariance matrix where its entry at the ith row and jth column is $\mathbb{K}\left(X^{(1)}_{i,:}, X^{(2)}_{j,:}\right)$. To save the space, let $\mu_X = \mu(X)$ and $\mathbb{K}_{X^{(1)},X^{(2)}} = \mathbb{K}\left(X^{(1)}, X^{(2)}\right)$. In the regression models, the observational variance function Λ is additionally required. For any $N \times D$ matrix X, $\Lambda(X)$ is the $N \times N$ diagonal matrix where its nth diagonal entry is $\Lambda(X_n)$. Similarly, let $\Lambda_X = \Lambda(X)$.

1.4.2 Symbols

$\propto$	proportional to
$\sim$	distributed by
$\mathbb{R}$	the set of real numbers
$\mathbb{R}_+$	the set of nonnegative real numbers
$\vec{0}$	the zero vector
$\mathbb{I}$	the identity matrix
$v = v_{1:T}$	a T-dimensional (column) vector
v_i	the ith entry of a vector v

$v_{i:j}$	the truncated vector of a vector v by selecting the entries from i to j
$v^{\mathbb{T}}$	the transpose of a vector v
$diag\left(v\right)$	the diagonal matrix where the diagonal entries are from a vector v
$\mathrm{A} = \left[\mathrm{A}_{\mathrm{ij}}\right]_{\mathrm{i,j}=1}^{\mathrm{M,N}}$	a $\mathrm{M} \times \mathrm{N}$ matrix
A_{ij}	the entry from the ith row and jth column of a matrix A
$\mathrm{A}_{\mathrm{i},:}$	the ith row of a matrix A
$\mathrm{A}_{:,\mathrm{j}}$	the jth column of a matrix A
$\mathrm{A}^{\mathbb{T}}$	the transpose of a matrix A
A^{-1}	the inverse of a matrix A
$\left\|\mathrm{A}\right\|$	the determinant of a matrix A
$diag\left(\mathrm{A}\right)$	the (column) vector consisting of the diagonal entries of a matrix A
$p\left(y\middle\|\Theta\right)$	the conditional distribution of y given Θ
$\mathcal{N}\left(y\middle\|\mu,\Sigma\right)$	the (multivariate) normal probability density at y with a mean vector μ and covariance matrix Σ
$\mathcal{GP}\left(\beta\middle\|\mu,\mathbb{K}_{\mathrm{XX}}\right)$	the Gaussian process on β with an input X, mean function μ and covariance kernel function $\mathbb{K}$
$\Gamma\left(y\middle\|\alpha,\beta\right)$	the Gamma probability density at y with a shape parameter α and rate parameter β
$\mathcal{T}\left(y\middle\|\mu,\sigma^2,\nu\right)$	the generalized Student's t-distribution at y with a location parameter μ, scale parameter σ and degree of freedom ν
$\mathbb{E}_{p(\mathrm{X})}\left[f(\mathrm{X})\right]$	the expectation of $f(\mathrm{X})$ over the random variable X that follows a probability distribution p
$\mathbb{D}_{\mathrm{KL}}\left(q\middle\|p\right)$	the Kullback-Leibler divergence from p to q

1.4.3 Acronyms

AR	autoregressive
BIC	Bayesian information criterion
CO_2	carbon dioxide
$\delta^{11}B$	a variation derived from the boron isotopes ^{11}B and ^{10}B
$\delta^{18}O$	a measure of the ratio of stable isotopes ^{18}O and ^{16}O
DGP	deep Gaussian process
DNN	deep neural network
DTW	dynamic time warping
ELBO	evidence lower bound
EM	expectation-maximization
EP	expectation propagation
FITC	fully independent training conditional
GP	Gaussian process
GPC	Gaussian process classification
GPLVM	Gaussian process latent variable model
GPR	Gaussian process regression
GPST	Gaussian process state-space model
HMC	Hamiltonian Monte Carlo
HMM	hidden Markov model
KL	Kullback-Leibler
LOO-CV	leave-one-out cross-validation
MAP	maximum a posteriori
MCMC	Markov-chain Monte Carlo
ML	maximum likelihood
NN	neural network

NWR	Nadaraya-Watson kernel regression
ODE	ordinary differential equation
OU	Ornstein-Uhlenbeck
PPCA	probabilistic principal component analysis
RNN	recurrent neural network
SE	squared exponential
SSM	state-space model
SST	sea surface temperature
TEX_{86}	tetraether index of 86 carbon atoms
$U_{37}^{K'}$	unsaturation index of di- versus tri- unsaturated C_{37} alkenones
VFE	variational free energy

2

Benthic δ^{18}O Stack Construction

This chapter is based on the paper in preparation [Lee et al. (2019)] (The previous title is 'Dual Proxy Gaussian Process Stack: Integrating Benthic δ^{18}O and Radiocarbon Proxies for Inferring Ages on Ocean Sediment Cores'). Thus, some expressions and figures are the same with those in that paper. This work has been done with the collaboration of Dr. Lorraine Lisiecki, Devin Rand, Dr. Geoffrey Gebbie and my advisor, Dr. Charles Lawrence.

δ^{18}O is the ratio of stable isotopes ^{18}O and ^{16}O relative to the laboratory standard. It is used as a proxy for inferring the past water temperatures, salinity and global ice volumes in paleoceanography. The δ^{18}O of benthic foraminifera (benthic δ^{18}O) live on the seafloor and it is regarded as a global climate parameter because about half of its variance through time is explained by the

global ice volume [Spratt & Lisiecki (2016)] while deep-water temperature and salinity are relatively similar over space.

The characteristic that benthic $\delta^{18}O$ is a global parameter has been utilized for the alignment of ocean sediment cores. Once cores are aligned, direct age information obtained by the direct age proxies separately stored in each core is transferred from one to another, thus more complete and comprehensive age inferences to all aligned cores are yielded.

The multiple core alignment problem is not straightforwardly resolved by the pairwise alignment algorithms for the following reasons: 1) The number of pairs is quadratic to the number of cores thus considering all such pairwise alignments becomes intractable as the number of cores increases, and 2) Pairwise alignments do not guarantee the consistency of the multiple alignments, i.e., the pairwise alignment between A and C indirectly obtained by those of core A and B and of B and C would not be consistent to the direct pairwise alignment of A and C.

Inspired by the profile hidden Markov models (HMMs) in computational biology [Durbin et al. (1998)], iterating the step of aligning each core to a profile (or a reference in more common terminology) and that of constructing (updating) the profile from the alignments is an alternative solution for the multiple core alignment problem. For benthic $\delta^{18}O$, the term 'stack' [Lisiecki & Raymo (2005)] is instead used as 'profile' that is more prevalent in computational biology.

Thus, the multiple core alignment with stacks is involved with the following two algorithms: the core alignment algorithm and the stack construction algorithm. In this section, we design a continuous core alignment algorithm that simultaneously utilizes ^{14}C as the direct proxy and benthic $\delta^{18}O$ as the indirect one and a continuous stack construction algorithm based on the homoscedastic Gaussian process regression (GPR) in section B.2.1.

2.1 Related Work

For the direct age inference, radiocarbon (^{14}C) is frequently used as a proxy by allowing to access calendar ages associated to the sampled core depth based on the age models and calibration curves [Reimer et al. (2013); Hogg et al. (2013)]. [Christen & Sergio (2009)] designed an algo-

rithm for constructing sediment core age models from ^{14}C data based on a generalized Student's t-distribution: it addresses the uncertainties from ^{14}C measurements and reservoir effects.

Moreover, various studies have been developing dynamic models that also reflect the varying sediment accumulation rates over time. [Blaauw & Christen (2005)] adopt a piecewise linear approach with automatic section selection and impose constraints on accumulation rates. [Haslett & Parnell (2008)] choose a bivariate monotone Markov process with gamma increments. [Blaauw & Christen (2011)] develop a software which is called Bacon (Bayesian Accumulation) that adopts an autoregressive gamma process for accumulation rates, a generalized Student's t-distribution for radiocarbon data and an adaptive Markov-chain Monte Carlo algorithm for sampling ages.

The benthic $\delta^{18}O$ utilization for indirect age inferences has the following history. [Lisiecki & Lisiecki (2002)] developed a deterministic alignment algorithm, Match, that uses dynamic programming. [Lisiecki & Raymo (2005)] ran Match on benthic $\delta^{18}O$ data from 57 globally distributed deep-sea sediment cores for aligning them and calculating a stack, LR04: it is commonly used as a standard reference for benthic $\delta^{18}O$ change over the past 5.3 Myr.

[Lin et al. (2014)] designed an algorithm HMM-Match based on a probabilistic alignment model to the stack. Its emission model for $\delta^{18}O$ data is based on Gaussian distribution with time varying mean $\delta^{18}O$ values from LR04 and a constant core-dependent standard deviation learned by the Baum-Welch algorithm. The transition model explains the probability distribution of accumulation rates based on a log-normal mixture learned by radiocarbon observations from 37 cores. [Ahn (2016)] constructed a new Prob-stack from 180 globally distributed benthic $\delta^{18}O$ records with an algorithm HMM-Stack that is based on HMM-Match. Prob-stack allows the change of variances over time.

HMM-based models cannot avoid assuming inferred ages on discrete spaces, thus ages must be discretized. This is problematic for the alignments when the cores have higher resolution than the target stack or when a proportional accumulation rate model is chosen as the above-mentioned algorithms. Increasing the resolution of the target stack becomes intractable since the time com-

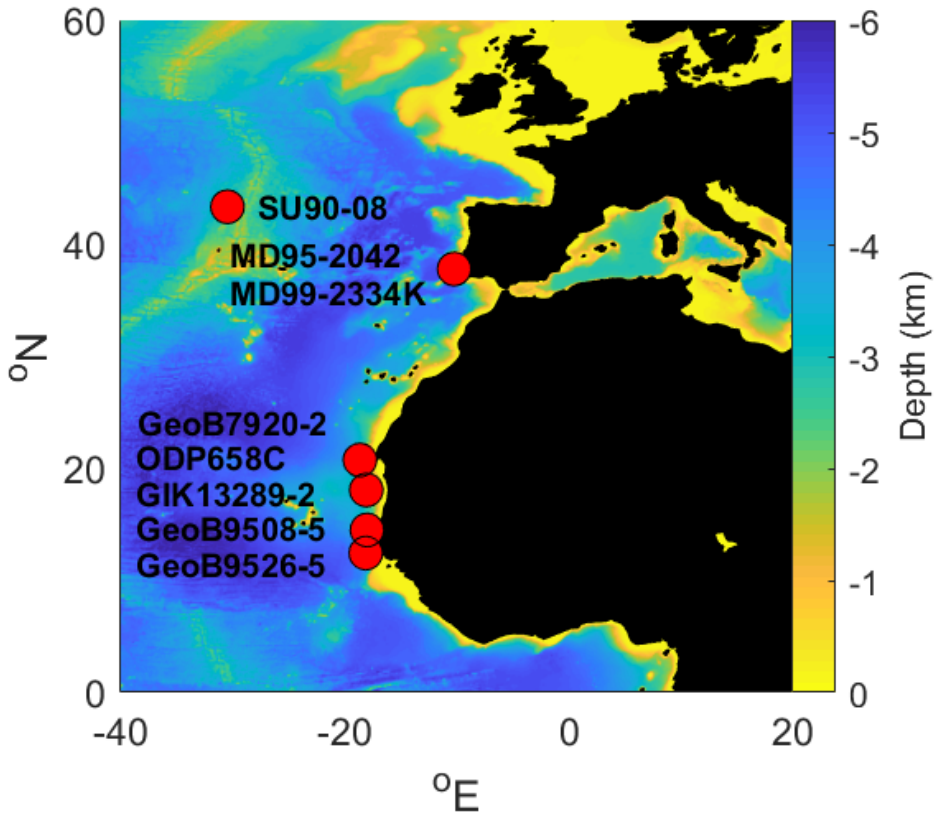
plexity of HMMs is quadratic to the size of hidden spaces. Stack construction algorithms based on the expectation-maximization (EM) algorithm [Dempster et al. (1977)] such as Baum-Welch (section A.1.3) also have a structural problem that stems from the relationship between the resolution and coverage of the cores: stack construction requires a number of cores that have δ^{18}O data enough to cover its query ages. Simple interpolation is not a good solution because it does not account for the increasing uncertainty from missing observations.

The following subsections demonstrate the core alignment algorithm and stack construction algorithm that are based on the Signal Alignment algorithm with Gaussian Process Regression profiles (SA-GPR) [Lee et al. (2019)]. Ages are sampled from continuous spaces in the alignment algorithm based on the particle smoother (section A.3.2) and Metropolis-Hastings (section A.4.2), thus the model does not suffer from the resolution-relevant drawbacks as the HMM-based ones. The stack is constructed from the sample-specific GPR models (section B.2.1), so the construction can be done regardless of the resolutions and number of cores.

2.2 Research Motive and Data

The model assumes that benthic δ^{18}O data in the input cores show the same pattern over ages: it is roughly true for all ocean sediment cores because benthic δ^{18}O is considered as a global parameter, but the local variability between core locations might cause discordance in millennial scale events. [Skinner & Shackleton (2005); Stern & Lisiecki (2014)] report that local effects could result in age model errors up to 4 kiloyears. Such local variability possibly stems from the asynchronous temperature changes and ocean circulation rates by ice melt after the Last Glacial Maximum (19-23 kiloyears ago) [Adkins et al. (2002); McManus et al. (2004); Waelbroeck et al. (2011); Gebbie & Huybers (2012)].

The purpose of our research is for capturing such local variability by constructing a stack for each set of records that share the same local variability: we define such a set to be homogeneous and the associated stack is called a local stack. To validate our stack construction method, we ran six ocean sediment cores, GeoB7920-2, GeoB9508-5, GeoB9526-5, MD95-2042, MD99-2334



Core	Latitude	Longitude	Depth (m)	Citation
GeoB7920-2	20.75	−18.58	2278	Tjallingii et al. (2008), Collins et al. (2011)
GeoB9508-5	14.5	−17.95	2384	Mulitza et al. (2008)
GeoB9526-5	12.44	−18.06	3223	Zarriess & Mackensen (2011), Zarriess et al. (2011)
GIK13289-2	18.07	−18.01	2485	Sarnthein et al. (1994)
MD95-2042	37.8	−10.17	3146	Bard et al. (2004), Shackleton et al. (2000), Shackleton et al. (2004)
MD99-2334	37.8	−10.17	3146	Skinner et al. (2003), Skinner & Shackleton (2004)
ODP658C	20.75	−18.58	2273	Demenocal et al. (2000)
SU90-08	43.35	−30.41	3080	Paterne et al. (1999), Grousset (1993)

Figure 2.1: Locations of cores GeoB7920-2, GeoB9508-5, GeoB9526-5, MD95-2042, MD99-2334, ODP658C, GIK13289-2 and SU90-08.

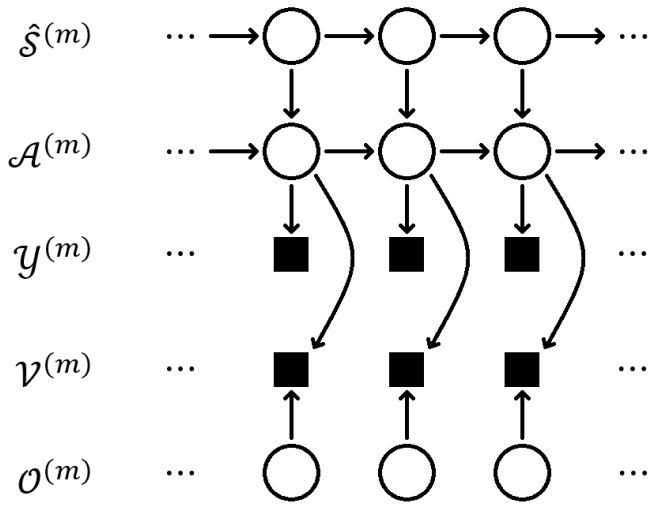


Figure 2.2: A graphical representation of our core alignment model.

and ODP658C, to construct a dual proxy local stack, Deep Northeastern Atlantic (DNEA) stack. We then use this local stack to infer dual proxy ages of two records which are homogeneous but not used in constructing the local stack, GIK13289-2 and SU90-08. Details of the data are given in figure 2.1.

2.3 Core Alignment Algorithm

The core alignment algorithm aims at inferring ages at the query depths in the input cores: i.e., cores to be aligned are synchronized by their age assignments. Our probabilistic model does not return the ages as their estimates but samples, thus the results are given as a set of sampled age paths for each input core. Before giving the detailed demonstration, the following notations are defined:

- $\mathcal{D}^{(m)} = \left\{d_n^{(m)}\right\}_{n=1}^{\mathrm{N}_m}$: depths of core m, shallower to deeper, at which the observations are made.
- $\mathcal{A}^{(m)} = \left\{\mathrm{A}_n^{(m)}\right\}_{n=1}^{\mathrm{B}_m}$: hidden ages of core m that are assigned to $\mathcal{D}^{(m)}$.
- $\mathcal{V}^{(m)} = \left\{v_n^{(m)}\right\}_{n=1}^{\mathrm{N}_m}$: benthic $\delta^{18}\mathrm{O}$ observations at $\mathcal{D}^{(m)}$ of core m. If no value is assigned,

leave it $\emptyset$.

- $\mathcal{Y}^{(m)} = \left\{y_n^{(m)}\right\}_{n=1}^{\mathrm{N}_m}$: ^{14}C measurements at $\mathcal{D}^{(m)}$ of core m. If no value is assigned, leave it $\emptyset$.

The accumulation (sedimentation) rate model controls the ages given depths. Once ages are given to the depths, the models that associated benthic δ^{18}O observations or ^{14}C measurements follow are determined by the stack (for benthic δ^{18}O) or the calibration model (for ^{14}C). Because the vehicles that contains benthic δ^{18}O and ^{14}C are different, it does not go too far assuming that they are conditionally independent given ages. Thus, this core alignment problem can be interpreted as a state-space model. A graphical representation of our core alignment model is described in figure 2.2.

2.3.1 Transition Model

To define the transition model on the ages given depths, we utilize the dynamics of the accumulation rates. An accumulation rate from $d_{n+1}^{(m)}$ to $d_n^{(m)}$ is defined as follows:

$$\mathrm{S}_n^{(m)} \triangleq \frac{\mathrm{A}_{n+1}^{(m)} - \mathrm{A}_n^{(m)}}{r^{(m)}\left(d_{n+1}^{(m)} - d_n^{(m)}\right)} \tag{2.1}$$

, where $r^{(m)}$ is a depth-scale constant parameter for standardizing the accumulation rates.

While ages are still defined continuously, the accumulation rates are discretized into the following three states: $\{\mathbb{C}, \mathbb{A}, \mathbb{E}\}$ that stand for contraction, average and expansion, respectively. Let $\hat{\mathrm{S}}_n^{(m)}$ be the discretization of $\mathrm{S}_n^{(m)}$. Once $\hat{\mathrm{S}}_n^{(m)}$ is given, the transition model on the age is determined as follows:

$$p\left(\mathrm{A}_n^{(m)} \middle| \mathrm{A}_{n+1}^{(m)}, \hat{\mathrm{S}}_n^{(m)}\right) \propto \Gamma\left(\frac{\mathrm{A}_{n+1}^{(m)} - \mathrm{A}_n^{(m)}}{r^{(m)}\left(d_{n+1}^{(m)} - d_n^{(m)}\right)} \middle| \alpha, \beta\right) \cdot 1_{\left\{\frac{\mathrm{A}_{n+1}^{(m)} - \mathrm{A}_n^{(m)}}{r^{(m)}\left(d_{n+1}^{(m)} - d_n^{(m)}\right)} \in \mathrm{I}_{\hat{\mathrm{S}}_n^{(m)}}\right\}} \tag{2.2}$$

, where $\mathrm{I}_{\mathbb{C}} \triangleq (m, 0.9220)$, $\mathrm{I}_{\mathbb{A}} \triangleq [0.9220, 1.0850)$ and $\mathrm{I}_{\mathbb{E}} \triangleq [1.0850, \mathrm{M})$ for some thresholds

$0.9220 > m \geq 0$ and $\mathrm{M} > 1.0850$ to prevent extreme events.

We also give a dependency on the discretized accumulation rates for a state-transition matrix φ as follows:

$$p\left(\hat{\mathrm{S}}_n^{(m)} \middle| \hat{\mathrm{S}}_{n+1}^{(m)}\right) \triangleq \varphi^{(m)}_{\hat{\mathrm{S}}_{n+1}^{(m)}, \hat{\mathrm{S}}_n^{(m)}} \tag{2.3}$$

From (2.2) and (2.3), we have the following transition model:

$$p_t\left(\mathrm{A}_n^{(m)}, \hat{\mathrm{S}}_n^{(m)} \middle| \mathrm{A}_{n+1}^{(m)}, \hat{\mathrm{S}}_{n+1}^{(m)}\right) \triangleq p\left(\mathrm{A}_n^{(m)} \middle| \mathrm{A}_{n+1}^{(m)}, \hat{\mathrm{S}}_n^{(m)}\right) p\left(\hat{\mathrm{S}}_n^{(m)} \middle| \hat{\mathrm{S}}_{n+1}^{(m)}\right) \tag{2.4}$$

Thus, our transition model is of two layers and not continuous. Also, note that it is defined backwards: it follows the way how layers are accumulated – from deeper to shallower. By definitions, the transition model guarantees the monotonicity of ages: the deeper the layer is, the older it is. The shape and rate hyperparameters α and β of the gamma distribution in (2.2) are fixed in learning to avoid overfitting: their values are learned a priori from the same $^{14}\mathrm{C}$ dataset for learning the log-normal mixture model in [Lin et al. (2014)]. A gamma distribution fits into the data better than the log-normal mixture and shows thicker tails between the thresholds, thus more robust on the semi-extreme changes in the accumulation rates.

2.3.2 Emission Model

Here we specify the emission model $p_e\left(y_n^{(m)}, y_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right) = p\left(y_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right) p\left(y_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right)$ by defining each observational model $p\left(y_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right)$ and $p\left(y_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right)$.

For benthic $\delta^{18}\mathrm{O}$, we basically accept the assumption that each $\delta^{18}\mathrm{O}$ observation follows a Gaussian distribution with the mean μ_{S} and variance σ_{S}^2 defined in the target stack at the given age, as [Ahn (2016)] does. However, this assumption is not robust on the real data that might be contaminated by the noninformative outliers. To reflect it we also add another set of the hidden indicator states, $\mathcal{O}^{(m)} = \left\{\mathrm{O}_n^{(m)}\right\}_{n=1}^{\mathrm{N}_m}$, that determines whether or not the associated $y_n^{(m)}$ is an

outlier.

$$
\begin{aligned}
p\left(\mathrm{O}_n^{(m)}\right) &\triangleq \text{Bernoulli}\left(\mathrm{O}_n^{(m)} \middle| 0.05\right) \\
p\left(v_n^{(m)} \middle| \mathrm{A}_n^{(m)}, \mathrm{O}_n^{(m)} = 0\right) &= \mathcal{N}\left(v_n^{(m)} \middle| \mu_{\mathrm{S}}\left(\mathrm{A}_n^{(m)}\right) + h^{(m)}, \sigma_{\mathrm{S}}^2\left(\mathrm{A}_n^{(m)}\right)\right) \\
p\left(v_n^{(m)} \middle| \mathrm{A}_n^{(m)}, \mathrm{O}_n^{(m)} = 1\right) &= \frac{1}{2}\sum_{k=1}^{2} \mathcal{N}\left(v_n^{(m)} \middle| \mu_{\mathrm{S}}\left(\mathrm{A}_n^{(m)}\right) + 3\left(-1\right)^k \sigma_{\mathrm{S}}\left(\mathrm{A}_n^{(m)}\right) + h^{(m)}, \sigma_{\mathrm{S}}^2\left(\mathrm{A}_n^{(m)}\right)\right)
\end{aligned}
\tag{2.5}
$$

, where0 $h^{(m)}$ is a core-specific shift parameter that synchronizes the δ18O observations vertically. As (2.5), $O_n^{(m)} = 1$ if and only if $y_n^{(m)}$ is an outlier.

From (2.5), $p\left(v_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right)$ is obtained by marginalizing $\mathrm{O}_n^{(m)}$ as follows:

$$
p\left(v_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right) = 0.05p\left(v_n^{(m)} \middle| \mathrm{A}_n^{(m)}, \mathrm{O}_n^{(m)} = 1\right) + 0.95p\left(v_n^{(m)} \middle| \mathrm{A}_n^{(m)}, \mathrm{O}_n^{(m)} = 0\right) \tag{2.6}
$$

For ^{14}C, we adopt the generalized Student's t-distribution in [Christen & Sergio (2009)] with the calibration curve that consists of the mean μ_{C} and variance σ_{C}^2 defined over the calendar ages:

$$
p\left(y_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right) = \mathcal{T}\left(y_n^{(m)} \middle| \mu_{\mathrm{C}}\left(\mathrm{A}_n^{(m)}\right) + \varrho_n^{(m)}, \frac{b}{a}\left(\sigma_{\mathrm{C}}^2\left(\mathrm{A}_n^{(m)}\right) + \varsigma_n^{(m)}\right), 2a\right) \tag{2.7}
$$

, where $\varrho_n^{(m)}$ and $\varsigma_n^{(m)}$ are the input variables (assumed to be fixed) that are given with $y_n^{(m)}$ a priori, and $a = 3$ and $b = 4$ are fixed hyperparameters to control the model variability: we assume the same values that [Christen & Sergio (2009)] suggested.

Note that the above equalities hold if and only if $v_n^{(m)}$ and $y_n^{(m)}$ are assigned by the real values: if $\emptyset$ is assigned instead, then the values are identical to 1, which implies "no information".

The reason why we do not model $p\left(v_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right)$ based on the Student's t-distribution just as $p\left(y_n^{(m)} \middle| \mathrm{A}_n^{(m)}\right)$ is because eventually the outliers in the sampled age paths are discarded before the stack construction algorithm. Note that (2.5) derives the posterior $p\left(O_n^{(m)} \middle| \mathrm{A}_n^{(m)}, v_n^{(m)}\right)$ once both $\mathrm{A}_n^{(m)}$ (as a sample) and $v_n^{(m)}$ are given.

2.3.3 Details on the Inference

Now we are prepared: a state-space model is fully specified by its transition and emission models. However, model formulation does not tell anything about the tractable algorithm for inference. Here, we adopt a hybrid of the particle smoother (section A.3.2) for initializing sampled age paths and Metropolis-Hastings algorithm (section A.4.2) for refining those paths. Because each $A_n^{(m)}$ is one-dimensional, the proposal distribution can be localized: i.e., after running a few iterations, fix the region at each depth and draw particles in those regions only. Parameters are learned by the EM algorithm from the sampled age paths. Algorithm 1 summarizes the inference procedure of our core alignment algorithm.

Algorithm 1: Core Alignment Algorithm

```
initialize parameters
for each core do
    while converging do
        run particle smoother on the data to draw sampled age paths
        run Metropolis-Hastings algorithm on the data with the sampled age paths
        estimate parameters given the sampled and refined age paths
    end
    for each sampled and refined age path do
        estimate outliers
    end
end
return sampled and refined age paths with their outliers and parameters
```

A natural question arises as follows: why do we not use (mixture) Kalman smoother against all the odds of the particle smoother described in section A.3.2? Ages are supposed not to be reversible and the accumulation rate model is a bit skewed, thus any Gaussian model cannot replace the gamma distribution in (2.2). The following doubt may stem from our choice of Metropolis-Hastings: section A.4.3 suggests the Hamiltonian Monte Carlo (HMC) as a better MCMC algorithm. Because the transition model is not differentiable, it is not possible to apply HMC here.

Dynamic time warping (DTW) [Munich & Perona (1999); Petitjean et al. (2011, 2014)] is frequently used for the alignment problems, but not adopted here because of the following reasons:

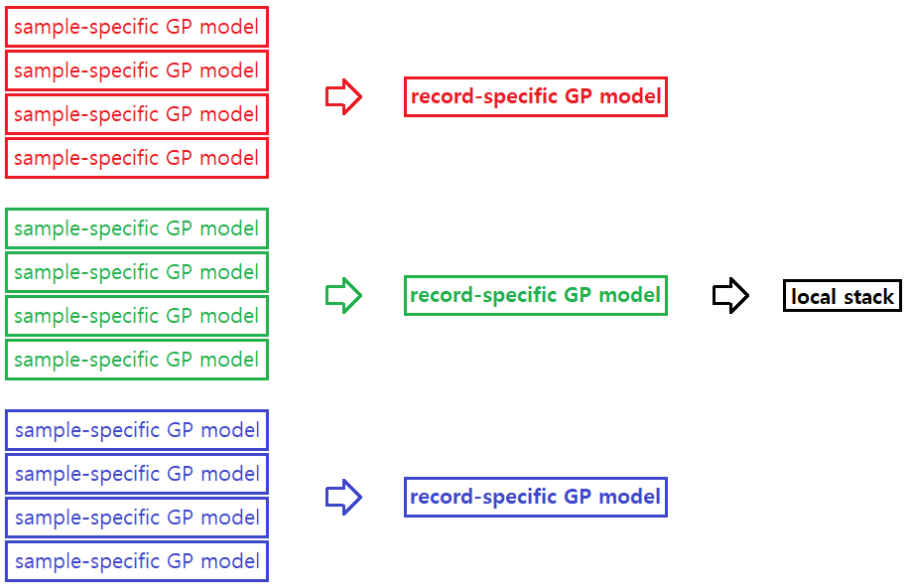


Figure 2.3: Three steps for the stack construction algorithm.

1) DTW has trouble in reflecting the dynamics of the latent series properly if the profile is not of high resolution, 2) results can be strongly dependent on the user-specific pairs and are often forced to match outputs that are apparently asynchronous, 3) the performance is sensitive to the penalties and loss functions in the transition and emission models, 4) the method is deterministic and does not reflect alignment uncertainty (which is beneficial if there are sub-optimal paths) and 5) DTW has a difficulty in addressing signals with heteroscedastic noises by the limitations on the loss function which is defined on the outputs directly.

We also sample outliers *after* drawing sampled and refined age paths that uses the marginalized observational model for benthic δ^{18}O: this is called the Rao-Blackwellized sampling that efficiently reduces the sampling variance of the estimation.

2.4 Stack Construction Algorithm

In this section, we assume to have obtained a set of sampled age paths for each core (record) by running the hybrid algorithm in section 2.3. The stack construction algorithm consists of the following three steps: 1) constructing sample-specific models, 2) averaging sample-specific models to obtain record-specific models and 3) getting the local stack from the record-specific

models. Figure 2.3 visualizes these steps.

2.4.1 Sample-specific Models

This step focuses on each sampled and refined age path (the kth one) of the core m with the associated benthic δ^{18}O observations $\left(\widetilde{\mathcal{A}}^{(m,k)}, \widetilde{\mathcal{V}}^{(m,k)}\right)$ after classifying and discarding outliers and synchronizing the benthic δ^{18}O observations vertically by the core-specific shift parameter $h^{(m)}$. Firstly, we applied homoscedastic Gaussian process regression (GPR) in section B.2.1 to the pair by regarding $\widetilde{\mathcal{A}}^{(m,k)}$ and $\widetilde{\mathcal{V}}^{(m,k)}$ as inputs and output, respectively, and derived the following posterior predictive model as the sample-specific one:

$$p\left(v\middle|a, \widetilde{\mathcal{A}}^{(m,k)}, \mathcal{V}^{(m)}\right) = p\left(v\middle|a, \widetilde{\mathcal{A}}^{(m,k)}, \widetilde{\mathcal{V}}^{(m,k)}\right) \triangleq \mathcal{N}\left(v\middle|\mu_{m,k}(a), \sigma^2_{m,k}(a)\right) \tag{2.8}$$

The Ornstein-Uhlenbeck (OU) kernel was selected, and kernel hyperparameters are tuned and shared by the sampled age paths of the same core. We adopted the exact GPR model without variational extension for dealing with large dataset because each core has less than 1000 benthic δ^{18}O observations thus such extension was not needed.

2.4.2 Record-specific Models

A record-specific model is designed to represent the corresponding record. Because ages are latent thus random, we define the record-specific model as the marginalized posterior predictive model that does not depend on the choice of ages. Because each $\widetilde{\mathcal{A}}^{(m,k)}$ is an independent and identically distributed samples drawn from $p\left(\mathcal{A}\middle|\mathcal{V}^{(m)}\right)$ by the core alignment algorithm in section 2.3, the following approximation is obtained:

$$\begin{aligned} p\left(v\middle|a, \mathcal{V}^{(m)}\right) &= \int p\left(v\middle|a, \mathcal{A}, \mathcal{V}^{(m)}\right) p\left(\mathcal{A}\middle|\mathcal{V}^{(m)}\right) d\mathcal{A} \\ &\approx \frac{1}{\mathrm{K}_m} \sum_{k=1}^{\mathrm{K}_m} p\left(v\middle|a, \widetilde{\mathcal{A}}^{(m,k)}, \mathcal{V}^{(m)}\right) = \frac{1}{\mathrm{K}_m} \sum_{k=1}^{\mathrm{K}_m} \mathcal{N}\left(v\middle|\mu_{m,k}(a), \sigma^2_{m,k}(a)\right) \end{aligned} \tag{2.9}$$

If there is only one record m in the homogeneous set, $p\left(v\middle|a,\mathcal{V}^{(m)}\right)$ is then supposed to be a local stack that represents the set. Because we want to define the local stack as a Gaussian model, by the model reduction method based on the moment-matching [Murphy (2012)], we have the following Gaussian approximation (2.10) for the Gaussian mixture (2.9) as the record-specific model:

$$\begin{aligned} p\left(v\middle|a,\mathcal{V}^{(m)}\right) &\approx \frac{1}{\mathrm{K}_m}\sum_{k=1}^{\mathrm{K}_m}\mathcal{N}\left(v\middle|\mu_{m,k}\left(a\right),\sigma_{m,k}^2\left(a\right)\right) \approx \mathcal{N}\left(v\middle|\mu_m\left(a\right),\sigma_m^2\left(a\right)\right) \\ \mu_m\left(a\right) &\triangleq \frac{1}{\mathrm{K}_m}\sum_{k=1}^{\mathrm{K}_m}\mu_{m,k}\left(a\right) \\ \sigma_m^2\left(a\right) &\triangleq \frac{1}{\mathrm{K}_m}\sum_{k=1}^{\mathrm{K}_m}\left(\sigma_{m,k}^2\left(a\right)+\left(\mu_{m,k}\left(a\right)-\mu_m\left(a\right)\right)^2\right) \end{aligned} \tag{2.10}$$

2.4.3 Local Stack

Now we have a set of record-specific models obtained in the previous section. Let us recall the assumption that the input cores are homogeneous, i.e., they share the same local variability of the global parameter benthic $\delta^{18}\mathrm{O}$ thus are supposed to follow the same Gaussian model, a local stack:

$$p\left(v|a,\mathcal{V}\right)=\mathcal{N}\left(v\middle|\mu_{\mathrm{S}}\left(a\right),\sigma_{\mathrm{S}}^2\left(a\right)\right) \tag{2.11}$$

Let us assume that we have drawn M independent and identically distributed observations $\{v_m\}_{m=1}^{\mathrm{M}}$ from $p\left(v|a,\mathcal{V}\right)$. Here, our variational method assumes that the joint distribution of $(v_1,v_2,\cdots,v_{\mathrm{M}})$ is approximated by the product of the record-specific models $p\left(v\middle|a,\mathcal{V}^{(m)}\right)$, i.e., $\{v_m\}_{m=1}^{\mathrm{M}}$ is also considered as a set of samples where each v_m was drawn from the corresponding

$p\left(v|a,\mathcal{V}^{(m)}\right)$, respectively. The objective function to minimize is given as follows:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}\left(p\left(\cdot|a,\mathcal{V}\right)^{\mathrm{M}}\middle|\middle|\prod_{m=1}^{\mathrm{M}}p\left(\cdot|a,\mathcal{V}^{(m)}\right)\right)=\sum_{m=1}^{\mathrm{M}}\mathbb{D}_{\mathrm{KL}}\left(p\left(\cdot|a,\mathcal{V}\right)||p\left(\cdot|a,\mathcal{V}^{(m)}\right)\right)\\
&=\sum_{m=1}^{\mathrm{M}}\mathbb{D}_{\mathrm{KL}}\left(\mathcal{N}\left(\cdot|\mu_{S}\left(a\right),\sigma_{S}^{2}\left(a\right)\right)||\mathcal{N}\left(\cdot|\mu_{m}\left(a\right),\sigma_{m}^{2}\left(a\right)\right)\right)\\
&=\frac{1}{2}\sum_{m=1}^{\mathrm{M}}\left(\frac{\sigma_{S}^{2}\left(a\right)}{\sigma_{m}^{2}\left(a\right)}+\frac{\left(\mu_{S}\left(a\right)-\mu_{m}\left(a\right)\right)^{2}}{\sigma_{m}^{2}\left(a\right)}-1+\log\frac{\sigma_{m}^{2}\left(a\right)}{\sigma_{S}^{2}\left(a\right)}\right)
\end{aligned}\tag{2.12}$$

The minimizer $\left(\mu_{\mathrm{S}}\left(a\right),\sigma_{\mathrm{S}}^{2}\left(a\right)\right)$ of the KL divergence (2.12) are given by the following closed forms:

$$\begin{aligned}
\mu_{S}\left(a\right)&=\sum_{m=1}^{\mathrm{M}}\frac{\mu_{m}\left(a\right)}{\sigma_{m}^{2}\left(a\right)}\Big/\sum_{m=1}^{\mathrm{M}}\frac{1}{\sigma_{m}^{2}\left(a\right)}\\
\sigma_{S}^{2}\left(a\right)&=\mathrm{M}\Big/\sum_{m=1}^{\mathrm{M}}\frac{1}{\sigma_{m}^{2}\left(a\right)}
\end{aligned}\tag{2.13}$$

Theoretically, one can analytically compute $\mu_{\mathrm{S}}\left(a\right)$ and $\sigma_{\mathrm{S}}^{2}\left(a\right)$ for an arbitrary query age a by computing the sample-specific models in (2.8), the record-specific models in (2.10) and finally the local stack in (2.13), thus the stack is continuously defined. In practice, however, computing the mean and variance in (2.13) for a fine discrete set of ages and interpolating them for arbitrary query inputs is more efficient than the theoretic method: note that (2.13) implies that the constructed stack has continuous mean and variance functions.

2.5 Results and Discussion

The steps described in section 2.3 and 2.4 present the core idea of SA-GPR [Lee et al. (2019)]: note that the same framework is working for any transition and emission models. Figure 2.4 shows a diagram of SA-GPR: here we use the words 'motif' instead of 'profile' or 'stack' and 'signal' instead of 'sequence' or 'core', as the more general terms, i.e., both core alignments and stacks are obtained by iterating the core alignment algorithm in section 2.3 and stack construction algorithm in section 2.4 until convergence.

We constructed a dual proxy (that means both direct and indirect proxies were utilized) local

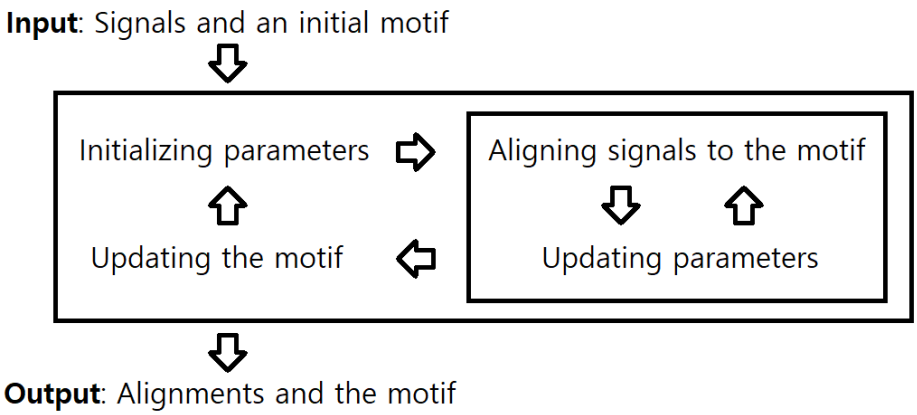


Figure 2.4: A simple diagram of the SA-GPR. Each box is iterated until convergence.

DNEA stack for the deep northeast Atlantic by assuming that two cores from the Iberian margin (MD95-2042 and MD99-2334) and four cores from the northwest African continental slope (GeoB7920-2, GeoB9508-5, GeoB9526-5 and ODP658C) are synchronous enough to be considered homogeneous. The availability of ^{14}C allows direct access to the calendar ages to enhance the age inferences: this is especially useful in the Holocene (after 11.7 ka) where the $\delta^{18}O$ signal-to-noise is low. We used the regional Deep North Atlantic (DNA) stack from [Lisiecki & Stern (2016)] to initialize the iterative algorithm. The open-source software for the stack construction and core alignment algorithms is available at `http://github.com/eilion/DPGP-Stack`, runs on MATLAB.

The upper and middle panels of figure 2.5 shows the alignments of the six cores to DNA stack and to DNEA stack. Variability in benthic $\delta^{18}O$ in the DNA stack is considerably larger than that in the DNEA stack: note that most of the medians are inside the 1-sigma of the DNA stack, which is only supposed to contain 68% of them. Higher variances in the DNA stack compared to the DNEA stack may stem from benthic $\delta^{18}O$ differences within the broader north Atlantic region, record-specific mean shifts applied to the DNEA stack but not the DNA stack, and/or the discrete nature of the algorithm used to construct the DNA stack. Another contributing factor to the tighter DNEA stack variance is the automatic detection and removal of outliers. The tighter variance will contribute to less uncertainty in ages inferred from this stack.

The smoothness of the local stack stems from the fact that the GP model captures correlations

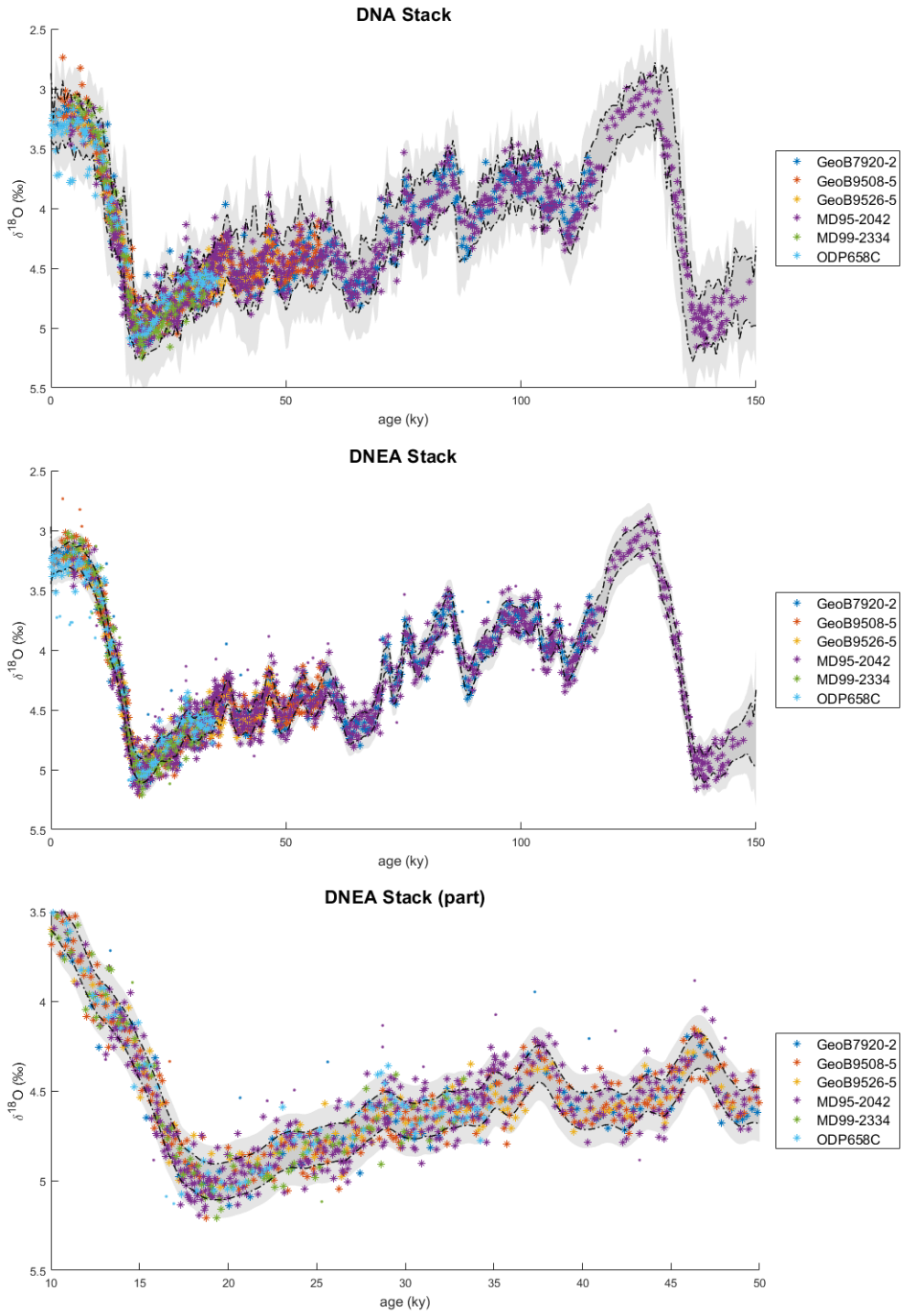


Figure 2.5: Core alignments to stacks. The upper panel (a) shows those to the DNA stack, while the middle panel (b) is our dual proxy local DNEA stack. Stars indicate medians of δ^{18}O data classified as inliers and dots represent outliers, after shifting them vertically by the core-specific shift parameters. The darker and brighter gray regions are the 1-sigma and 2-sigma of the stacks, respectively. The dot-dash line indicates their boundary. The lower panel (c) is a portion of the DNEA stack at 10-50 kiloyears.

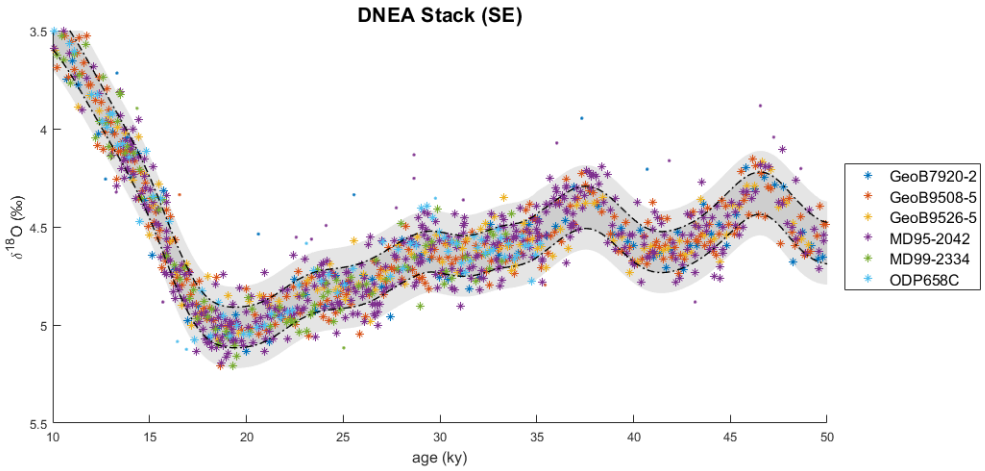


Figure 2.6: A part of the DNEA stack constructed with the SE kernel.

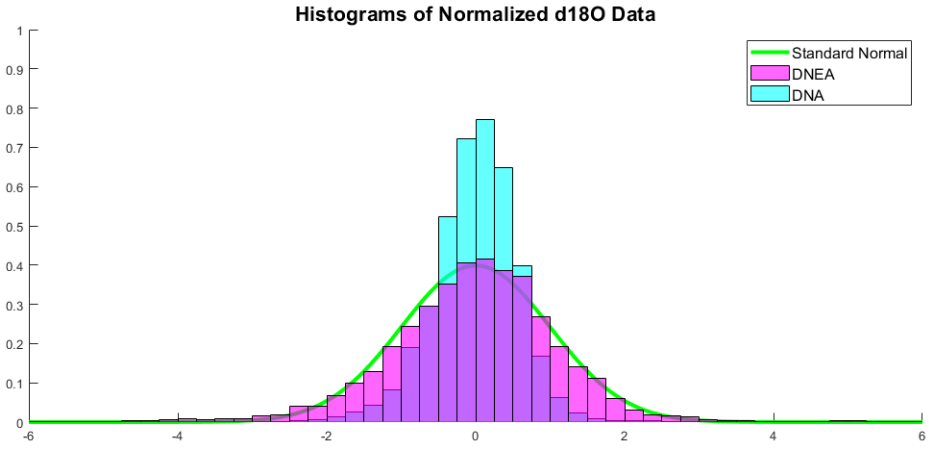


Figure 2.7: Histograms of δ^{18}O to our dual proxy DNEA stack. Each δ^{18}O is standardized by the mean and standard deviation of the stack at the median of its sampled ages.

between all the data points with a heavier weight placed on near neighbors, thus limiting sudden large changes. Although our dual proxy stack is smoother than the DNA stack, it still captures well-known millennial-scale climate events. For example, the bottom panel of figure 2.5 shows four peaks at 24, 29, 38 and 46 kiloyears, which correspond to the Heinrich events H2 to H5 [Skinner et al. (2013); Lisiecki & Stern (2016)]. The ability to resolve such short-lived features will improve the accuracy of age estimates for cores with high-resolution δ^{18}O records. We also tried the squared-exponential (SE) kernel but the results failed to reflect H2 and H3 at 24 and 29 kiloyears respectively, as figure 2.6.

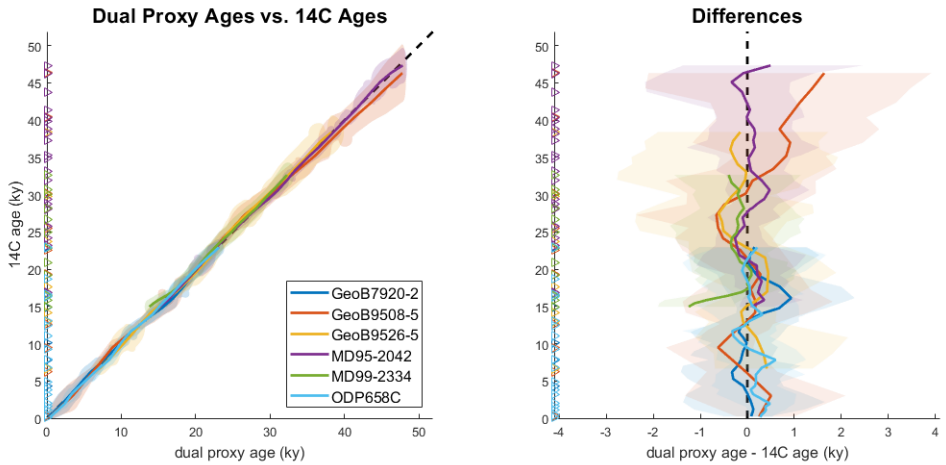


Figure 2.8: The comparison between inferred dual proxy ages and ^{14}C ages of records. Shaded regions indicate 95% confidence regions and the black dashed line is just a diagonal.

Figure 2.7 shows the histograms of standardized $\delta^{18}O$ from the six cores to the DNA and DNEA stacks, respectively. The fact that there is only a small departure from the standard normal distribution to the DNEA stack supports the validity of the Gaussian assumption. Figure 2.8 compares inferred ages by using both benthic $\delta^{18}O$ and ^{14}C (dual proxy) to those using ^{14}C only. Overall agreement between cores (to within uncertainty) supports our assumption that benthic $\delta^{18}O$ is synchronous and homogeneous among sites included in constructing DNEA stack. Some departures from the diagonal are expected, considering influences from their $\delta^{18}O$ data.

We also present the alignments of two ocean sediment cores from the same region, GIK13289-2 and SU90-08. We infer ages for these cores from their proxies and alignment to the DNEA stack based on the assumption that the aligned cores share the same local $\delta^{18}O$ effect as the stack. Figure 2.9 and 2.10 show the alignment results. The upper panels show that the shifted and aligned $\delta^{18}O$ data mainly fall within the stack's confidence intervals. The lower left panels show that the inferred ages mainly pass through the confidence intervals from individual ^{14}C proxies (SU90-08 has ^{14}C only between 10-40 kiloyears and beyond that ages are inferred only based on $\delta^{18}O$ alignment to the stack). Because stacks with dual proxy ages have narrower confidence intervals than single proxy stacks, they are more informative for stack alignment and produce smaller age uncertainties.

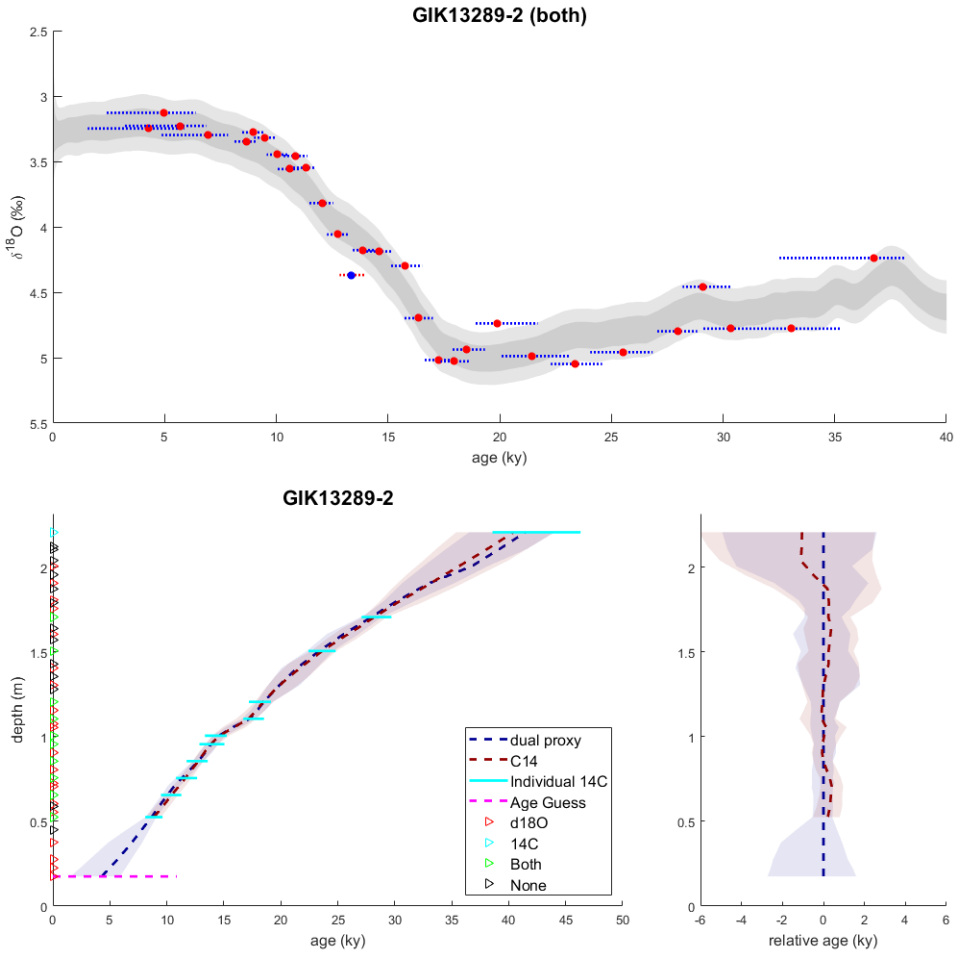


Figure 2.9: Dual proxy age inferences and alignments of GIK13289-2 to DNEA stack. In the upper figure, the darker and brighter areas show the 1-sigma and 2-sigma of the stacks, red points and blue dotted bars indicate medians and 95% confidence intervals of inlier age samples for $\partial^{18}\mathrm{O}$ data, blue points and red dotted bars are the outliers. In the left below figures, cyan bars indicate 95% confidence intervals obtained independently from $^{14}\mathrm{C}$ data, and blue and red areas show the 95% confidence bands of age inferences from $^{14}\mathrm{C}$ only and both proxies, respectively. In the right below figures, medians (dashed lines) and 95% confidence intervals of relative ages to the medians of dual proxy ages are shown.

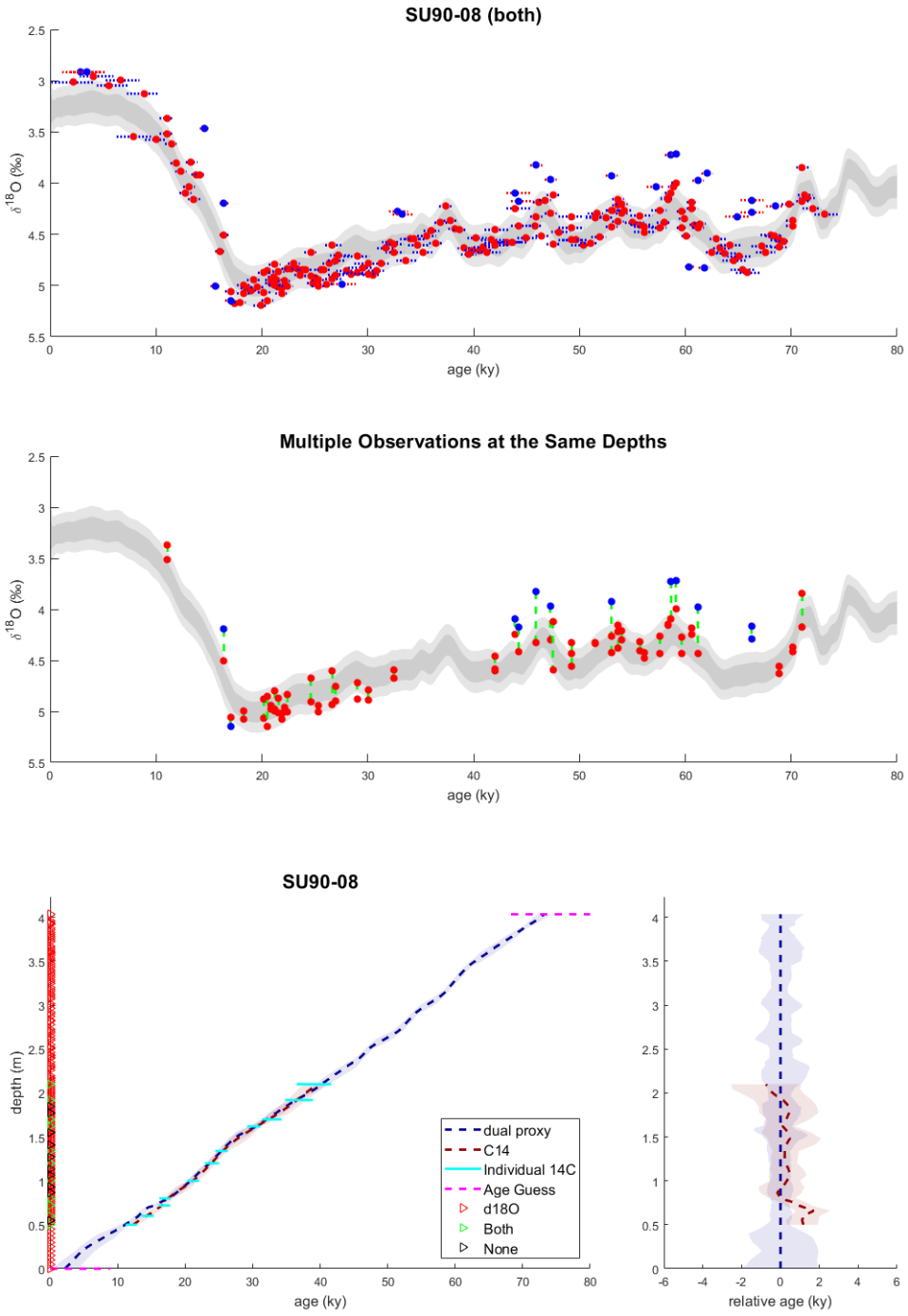


Figure 2.10: Dual proxy age inferences and alignments of SU90-08 to DNEA stack, similar to figure 2.9. This figure also shows how multiple observations at the same depths are dealt with in the alignment algorithm.

Our core alignment and stack construction algorithms share the advantages of the other stack-dependent alignment algorithms. Aligning cores does not only focus on finding the best matches between cores but also integrates the fragmented information stored in each core into the stack. Once the stack is constructed from a set of cores, it can be used as an alignment target for new cores.

Even if continuous sampling is available, it soon becomes insignificant without a continuously defined stack. Our stack construction algorithm depending on GPR is superior to the interpolation algorithms because it is theoretically clearer to exploit data to fill in unobserved outputs: the closer the query age is to the data, the higher the associated benthic $\delta^{18}O$ is correlated with them, based on the kernel covariance functions. Moreover, our method is nonparametric thus more robust on the general data.

There are several advantages of the new dual proxy method compared to using $\delta^{18}O$ or ^{14}C only. First, it gives a direct method for age inferences in a $\delta^{18}O$ stack. The dual proxy method can also be applied for other types of direct age constraints that can be expressed as distributions given their ages, such as dated magnetic reversals or volcanic ash layers. Second, ^{14}C determinations act as "anchor" which stabilizes the learning of transition parameters. Radiocarbon ages effectively protect the parameter learning procedure during the first EM iteration from ill-posed initial values. Third, the integration of both proxies should give more accurate age estimates by using more data. High resolution $\delta^{18}O$ measurements complement low-resolution ^{14}C data and substantial uncertainty bounds on ^{14}C-only inferences where ^{14}C data are sparse.

In addition, our stack construction algorithm can construct a stack with a limited number of cores because it uses all of the data from every core for inference of every age in the stack: previous methods such as HMM-Stack [Ahn (2016)] require a sufficient number of high-resolution records to estimate the mean and variance separately for each point in the stack. In addition, this dual proxy method returns means and variances in continuous time at arbitrary ages. Previous methods rely on less reliable interpolations between discrete ages.

While the state-transition matrix in section 2.3.1 can be learned either core-specifically or set-

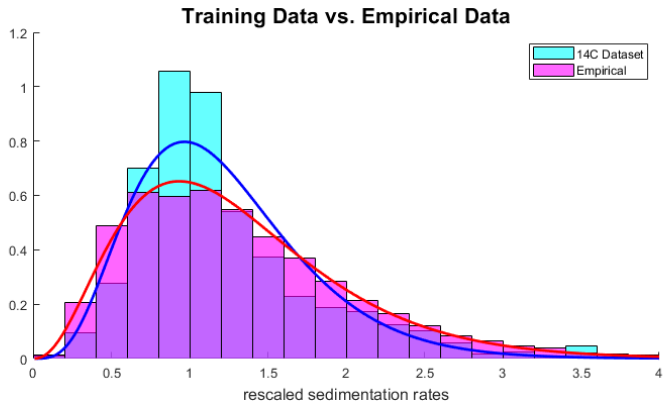


Figure 2.11: Histograms of the standardized accumulation rates. The blue histogram comes from the training ^{14}C dataset and the red one from the sampled age paths in constructing the DNEA stack. Thick lines indicate the inferred gamma distributions from the histograms, respectively.

specifically, the underlying accumulation model based on the gamma distribution is pre-trained and fixed. Gamma distributions tend to have thicker tails than log-normal mixtures learned from the same data, so alignments are more dependent on the emission model than the transition model, which makes the model more faithful to the observations. The comparison of our gamma distribution inferred from the training data and that from the empirical accumulation rates of cores in constructing the DNEA stack is in figure 2.11.

Here we avoid using the heteroscedastic GPR in section B.2.2 (instead of the homoscedastic GPR that we used) for regularization. However, if ∂^{18}O observations of the cores have higher resolution, this method will model more properly.

2.6 Future Work

As mentioned at the end of section 2.3.3, the reason why we did not consider HMC instead of Metropolis-Hastings is because the transition model is not differentiable. By giving two-layer hidden states, we could assume the second-order dependency of the ages given depth. If it is possible to alter this transition model to be differentiable, then the full model becomes differentiable with respect to the hidden states thus the core alignment algorithm will run faster by replacing the Metropolis-Hastings with HMC: note that the main bottleneck is the Metropolis-Hastings step. However, this cannot be done without the experts because sedimentation rate model itself

is geological thus must be consistent with the real data and theories certified by the community of paleoceanography.

The next step of local stack construction for each of the homogeneous sets one-by-one is, clearly, designing a more comprehensive construction algorithm that deal with multiple homogeneous sets simultaneously, i.e., considering how to allow homogeneous sets to communicate one another. With interdisciplinary collaborators, I am currently working on the hierarchical stack construction algorithm that automatically differentiate the dependency of δ^{18}O observations over cores with respect to their similarity. Here we cannot avoid considering variational extension because all data must be processed together while our stack construction algorithm in section 2.4 deviates the drawback of the GP models relevant to the size of data by constructing record-specific models and then merging them into a local stack, which depends on the assumption that the input cores are homogeneous.

Utilizing more age-relevant proxies, such as magnetic reversals or volcanic ash layers, is also an interesting topic. As discussed in section 2.5, the comprehensive framework of our method is flexible in processing age proxies, either direct or indirect. It could be utilized for constructing multiple stacks for indirect proxies other than δ^{18}O and utilizing more proxies at once gives an opportunity to exploit the more complete age information.

3

Alkenone/TEX_{86} Calibration Models

This section shows some extra works from the conference paper [Lee & Lawrence (2019)] that constructs a nonparametric alkenone calibration model based on the heteroscedastic Gaussian process regression.

The alkenone is a widely used proxy for inferring sea surface temperatures (SSTs) in paleoceanography and paleoclimatology. It is a long-chain unsaturated methyl and ethyl n-ketones produced by some phytoplankton species. Species that produce alkenone respond to environmental changes such as water temperatures by modifying the relative proportions of the saturated and unsaturated alkenones: more saturated alkenones are produced at higher temperatures. The

realization depends on the following indices U_{37}^{K} and $U_{37}^{K'}$ [Tierney & Tingley (2018)]:

$$\begin{aligned} U_{37}^{K} &= \frac{C_{37:2} - C_{37:4}}{C_{37:2} + C_{37:3} + C_{37:4}} \\ U_{37}^{K'} &= \frac{C_{37:2}}{C_{37:2} + C_{37:3}} \end{aligned} \tag{3.1}$$

, where $C_{37:2}$, $C_{37:3}$ and $C_{37:4}$ are the di-, tri- and tetra-unsaturated C_{37} alkenones. Though $C_{37:4}$ might be beneficial in prediction [Bard (2001); Bendle & Rosell-Mele (2004)], it does not exist in many marine sediment profiles, thus $U_{37}^{K'}$ is preferred.

The tetraether index of 86 carbons, TEX_{86} shortly, is another proxy that are utilized for reconstructing SSTs in paleoceanography and paleoclimatology. It is an index based on the relative cyclization of glycerol dialkyl glycerol tetraethers (GDGTs) that contain 0 to 3 cyclopentane moieties. The degree of cyclization is affected by the growth temperature [Wuchter et al. (2004)]. There are two ways of defining this index [Schouten et al. (2002); Kim et al. (2010)]:

$$\begin{aligned} TEX_{86} &= \frac{[GDGT-2] + [GDGT-3] + [Cren']}{[GDGT-1] + [GDGT-2] + [GDGT-3] + [Cren']} \\ TEX_{86}^{L} &= \frac{[GDGT-2]}{[GDGT-1] + [GDGT-2] + [GDGT-3]} \end{aligned} \tag{3.2}$$

, where Cren′ denotes the regioiomer of crenarchaeol that is a diagnostic biomarker for the Thaumarchaeota, the main producers of GDGTs [Sinninghe Damsté et al. (2002); Tierney & Tingley (2014)].

By definition, both $U_{37}^{K'}$ and TEX_{86} lie in the unit interval $[0, 1]$. Their underlying characteristics bring opportunities to utilize them for reconstructing past SSTs. A calibration model gives a relationship between the observations (known) and what we want to infer (unknown), either deterministic or probabilistic. Probabilistic calibration models aim at constructing distributions of the known given unknown, thus regarded as generative models (section 1.2.3): if additional information of the unknown independent of the known is available, then more reasonable inference on the unknown can be done in the framework of Bayesian statistics.

This chapter presents our ongoing work on the construction of probabilistic calibration mod-

els in various ways based on the Gaussian process regression (GPR) and their comparison. A novel multimodal GPR based on the Gibbs sampling (section A.4.1) and expectation-maximization (EM) algorithm [Dempster et al. (1977)] is also formulated in section 3.3.2.

3.1 Related Work

For the past years, both alkenone and TEX_{86} proxies have been gathered and calibrated, and thus consequently validated as effective proxies for SSTs. This subsection briefly introduces the previous works in modelling.

3.1.1 Alkenone Calibration Models

In response to the apparent linearity of the $U^{K'}_{37}$ observations over SSTs, several works have been focusing on constructing simple or region-specific linear regression models from their available data. For examples, [Prahl et al. (1988)] used 5 observations in range of 8 to 25 °C and resulted in $U^{K'}_{37} = 0.034 \cdot SST0.039$; [Müller et al. (1998)] obtained $U^{K'}_{37} = 0.033 \cdot SST + 0.044$ from 370 observations from 0 to 29 °C; [Sonzogni et al. (1997)] suggested $U^{K'}_{37} = 0.023 \cdot SST + 0.316$ for 40 observations from 24 to 29 °C in Indian ocean.

More recently, [Conte et al. (2006)] suspected the nonlinearity relationship above 26 °C and constructed a set of region-specific nonlinear models: $U^{K'}_{37} = -1.004 \cdot 10^{-4} \cdot SST^3 + 5.744 \cdot 10^{-3} \cdot SST^2 - 6.207 \cdot 10^{-2} \cdot SST + 0.407$ for 451 observations from Atlantic, $U^{K'}_{37} = 0.0391 \cdot SST - 0.1364$ for 131 observations from Pacific. Among all data sources including Indian and Southern Ocean, the global model for 629 observations is given as $U^{K'}_{37} = -5.256 \cdot 10^{-5} \cdot SST^3 + 2.884 \cdot 10^{-3} \cdot SST^2 - 8.4933 \cdot 10^{-3} \cdot SST + 9.898$.

[Tierney & Tingley (2018)] developed a Bayesian B-spline regression model, BAYSPLINE, with the treatment of data seasonality, from 1344 globally-gathered observations. The spline transformation is simply a linear combination of B-splines, so the regression model becomes a multiple linear regression. B-spline coefficients and model variance have priors and inferred in the Bayesian framework. The resulted model is linear below 23.4 °C with slope 0.034 that is

indistinguishable from [Prahl et al. (1988)], but the slope declines steadily to a minimum of 0.011 at the highest temperature 29.4 °C.

[Lee & Lawrence (2019)] constructed a GPR model as the calibration model from the same dataset of [Tierney & Tingley (2018)], after discarding those in the locations excluded from their analysis. The result of data with 1274 observations were used. After preprocessing on the remaining SST and $U_{37}^{K'}$ to be standardized, they applied the heteroscedastic GPR to them with a neural network kernel (section B.1.2) and constant mean prior function. Confirming the heteroscedasticity of data is one of their achievements.

3.1.2 TEX_{86} Calibration Models

Various simple models under certain conditions have been made so far. [Kim et al. (2008)] constructed a linear model $\mathrm{SST} = 56.2 \cdot \mathrm{TEX}_{86} - 10.78$ from 223 observations without data from Red Sea, those with $\mathrm{SST} < 5$ °C and those with residuals bigger than 1 sigma; [Liu et al. (2009)] resulted in $\mathrm{SST} = -16.332/\mathrm{TEX}_{86} + 50.475$ from 287 observations that include all their available data; [Kim et al. (2010)] gives two models: one is $\mathrm{SST} = 67.5 \cdot \log_{10} \mathrm{TEX}_{86}^{L} + 49.9$ without Red Sea data and the other is $\mathrm{SST} = 68.4 \cdot \log_{10} \mathrm{TEX}_{86} + 38.6$ without data from Red Sea and those with $\mathrm{SST} < 5$ °C.

As a more complicated model, [Tierney & Tingley (2014)] developed a Bayesian regression model, BAYSPAR, with the Gaussian process (GP) priors. Their region-specific linear models assume GP priors on the slope and intercept parameters over locations. They tested BAYSPAR under different conditions defined by either or some of Arctic data with $\mathrm{SST} < 3$ °C, Red Sea data, or Antarctic data with $\mathrm{SST} < 3$ °C were excluded. In either case, regional differences of slopes and intercepts were identified.

3.2 Research Motive and Data

The goal is to construct robust calibration models from the training data that consist of pairs of present (known) SSTs and either $U_{37}^{K'}$ or TEX_{86}, and then apply the models for inferring past

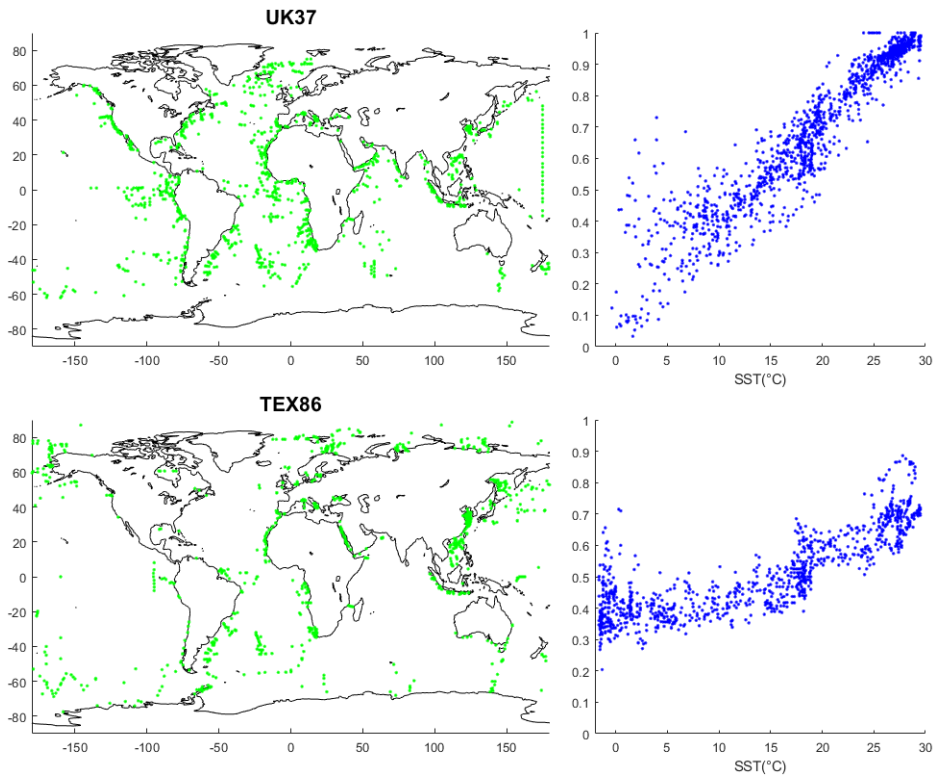


Figure 3.1: $U_{37}^{K'}$ and TEX_{86} data. Each left panel shows the data locations while right plots data values over SSTs.

SSTs given the ancient proxy observations. In more mathematical terminology, what we want to do is:

1. Constructing a calibration model $p\,(\text{Proxy}|\text{SST})$.

2. Applying it for $p\,(\text{SSTs}|\text{Proxies}) \propto p\,(\text{Proxies}|\text{SSTs})\,p\,(\text{SSTs}|\text{IK})$.

Here, IK stands for the independent knowledge of the SSTs from their locations or ages. Note that this simple framework could turn out to be a complicated probabilistic graphical model if $p\,(\text{SSTs}|\text{IK})$ forms a Markov chain (over ages), a spherical grid (over locations), or a more high-dimensional structure.

Figure 3.1 visualizes the dataset of $U_{37}^{K'}$ and TEX_{86}, respectively. We use the same dataset of [Tierney & Tingley (2018)] and [Tierney & Tingley (2014)]. The $U_{37}^{K'}$ dataset is comprised of [Arz et al. (2003); Bard et al. (2000); Barron et al. (2003); Barrows et al. (2007); Becker et al. (2015); Bendle & Rosell-Melé (2007); Cacho et al. (1999, 2001); Calvo et al. (2001, 2007); Caniupán et al. (2011); Chen et al. (2014); Conte et al. (2006); Benthien & Müller (2000); Chapman

et al. (1996); Doose et al. (1997); Dubois et al. (2009); Emeis et al. (2000); Fallet et al. (2012); Filippova et al. (2016); Fraser et al. (2014); Gutiérrez et al. (2011); Harada et al. (2006); Herbert et al. (1998, 2001); Ho et al. (2012); Horikawa et al. (2006); Huang et al. (1997); Ikehara et al. (1997); Jaeschke et al. (2007, 2017); Kaiser et al. (2005, 2008, 2014); Kennedy & Brassell (1992); Kienast & McKay (2001); Kienast et al. (2006, 2012); Kim et al. (2002a,b, 2003, 2004, 2007, 2015); Koutavas & Sachs (2008); Kudrass et al. (2001); Lamy et al. (2002); Leduc et al. (2007, 2010); Lee et al. (2001, 2008b,a); Leider et al. (2010); Liu & Herbert (2004); Lopes dos Santos et al. (2010); Lückge et al. (2009); Madureira et al. (1995); Marchal et al. (2002); Martrat et al. (2007); Max et al. (2012); McCaffrey et al. (1990); Mohtadi et al. (2011); Müller et al. (1998); Ohkouchi et al. (1999); Ostertag-Henning & Stax (2000); Pahnke & Sachs (2006); Pahnke et al. (2007); Pailler & Bard (2002); Pelejero & Grimalt (1997); Pelejero et al. (1999); Prahl et al. (1989, 2006, 2010); Rein et al. (2005); Rodrigo-Gámiz et al. (2015); Rodrigues et al. (2009, 2010); Romero et al. (2008); Rosell-Melé et al. (1995); Rosell-Melé (1998); Rosell-Melé et al. (2000); Rühlemann et al. (1999); Sachs (2007); Sawada et al. (1996); Sawada & Handa (1998); Schefuss et al. (2005); Schulz et al. (2002); Seki et al. (2004); Sepúlveda et al. (2009); Sikes et al. (1991); Sikes & Keigwin (1994, 1996); Sikes et al. (1997, 2009); Sonzogni et al. (1997, 1998); Sperling et al. (2003); Tao et al. (2012); Ternois et al. (2000); Tierney & Tingley (2018); Weldeab et al. (2007); Zhao et al. (1995); Zhao et al. (2006)], whereas the TEX_{86} dataset is consisting of [Castañeda et al. (2010); Chazen (2011); Chen et al. (2014); Fallet et al. (2012); Hernández-Sánchez et al. (2014); Ho et al. (2011, 2014); Hu et al. (2012); Jia et al. (2012); Kaiser et al. (2014); Kim et al. (2010); Leider et al. (2010); Lengger et al. (2014); Lü et al. (2014); Nieto-Moreno et al. (2013); Park et al. (2014); Richey et al. (2011); Seki et al. (2009, 2014); Shevenell et al. (2011); Shintani et al. (2011); Smith et al. (2013); Tierney & Tingley (2014); Trommer et al. (2009); Verleye (2011); Wei et al. (2011); Wu et al. (2012); Zell et al. (2014); Zhou et al. (2014)].

In either case, there are more than 1000 observations but less than 1500, thus the amounts of data are moderate. Though somewhat arbitrary linear regressions have been mostly applied so far, such relationships are not justified by theory, thus in fact there is no parametric model that not

only describes but also explains the relationship of SSTs and either $U_{37}^{K'}$ or TEX_{86}. As discussed in section 1.2.2, these characteristics fit well for the nonparametric approaches, such as GPR.

As shown in figure 3.1, both proxies show apparent heteroscedasticity over SSTs, thus it is reasonable that heteroscedastic GPR models tackle the problem. Though some of the previous research tend to either discard or adjust some observations based on their data-specific characteristics, here the full and raw dataset are used. To avoid the misspecification of GPR that assumes the monomodality of the data, we also apply a multimodal GPR model to the proxy datasets.

3.3 Models

As mentioned earlier, the GPR is a nonparametric regression model. However, it has a margin to reflect the parametric aspect in its structure: the mean prior function μ as discussed in Appendix B. For extending the previous works that are based on simple polynomial regressions (mostly linear), here we assume that μ is a cubic polynomial with unknown coefficient parameters to infer. Though the sizes of two datasets are between 1000 and 1500, a variational free energy (VFE) extension in section B.7 is applied in modelling. In this section, $\mathcal{X} = \{X_n\}_{n=1}^N$ and $\mathcal{Y} = \{Y_n\}_{n=1}^N$ are the sets of SSTs and proxy (either $U_{37}^{K'}$ or TEX_{86}) observations, respectively.

3.3.1 Homoscedastic and Heteroscedastic Gaussian Process Regression

Here we extend the heteroscedastic GPR model in [Lee & Lawrence (2019)]. Detailed derivations can be found in section B.2. In either proxy, the observational model is given as follows:

$$p(\mathcal{Y}|\beta, \mathcal{X}) = \mathcal{N}(\mathcal{Y}|\beta, \Lambda_{\mathcal{X}}) = \prod_{n=1}^{N} \mathcal{N}\left(Y_n|\beta_n, \Lambda_{X_n}\right) \tag{3.3}$$

, where β is a latent state that represents the regression of $\mathcal{Y}$ on $\mathcal{X}$. Recall that the GPR is homoscedastic if the variance function Λ is given as a constant function and heteroscedastic if not (section B.2). The prior on β is given as follows:

$$p(\beta|\mathcal{X}) = \mathcal{GP}(\beta|\varphi(\mathcal{X})\alpha, \mathbb{K}_{\mathcal{X}\mathcal{X}}) \tag{3.4}$$

, where $\mathbb{K}$ is a neural network kernel (section B.1.2) with erf as the activation function, $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^{\mathbb{T}} \in \mathbb{R}^4$ is a coefficient parameter and φ is a feature function defined as follows:

$$\varphi(\mathcal{X}) = \begin{pmatrix} 1 & X_1 & X_1^2 & X_1^3 \\ 1 & X_2 & X_2^2 & X_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_N & X_N^2 & X_N^3 \end{pmatrix} \tag{3.5}$$

An exact GPR is specified by (3.3) and (3.4). However, we also consider the VFE extension (section B.7) here: recall that it returns a closed-form posterior predictive model in regression. Let $\mathcal{X}_0$ be the set of pseudo-inputs. Then, the regression model is given as follows:

$$\begin{aligned} p(y|x, \mathcal{X}, \mathcal{Y}, \mathcal{X}_0) &= \mathcal{N}(y|\varphi(x)\alpha + m(x), \Lambda_x + s(x)) \\ m(x) &= \mathbb{K}_{x\mathcal{X}_0}\left(\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0} + \mathbb{K}_{\mathcal{X}_0\mathcal{X}}\Lambda_{\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}\mathcal{X}_0}\right)^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{X}}\Lambda_{\mathcal{X}}^{-1}(\mathcal{Y} - \varphi(\mathcal{X})\alpha) \\ s(x) &= \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}_0}\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0}^{-1}\mathbb{K}_{\mathcal{X}_0 x} + \mathbb{K}_{x\mathcal{X}_0}\left(\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0} + \mathbb{K}_{\mathcal{X}_0\mathcal{X}}\Lambda_{\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}\mathcal{X}_0}\right)^{-1}\mathbb{K}_{\mathcal{X}_0 x} \end{aligned} \tag{3.6}$$

Thus, the resulting model $p(y|x, \mathcal{X}, \mathcal{Y}, \mathcal{X}_0)$ in (3.6) can be understood as an extended polynomial regression: $\varphi(x)\alpha$ is the parametric regression of y and $m(x)$ is a nonparametric bias function, and the regression function is given by their sum $\varphi(x)\alpha + m(x)$.

Here, not only the kernel hyperparameters but also the coefficient α are unknown, so to be tuned and learned. The objective function is the evidence lower bound (ELBO) $\mathcal{L}$ given by the following closed form:

$$\begin{aligned} \log p(\mathcal{Y}|\mathcal{X}) \geq \mathcal{L} \triangleq \log \mathcal{N}\left(\mathcal{Y}\middle|\varphi(\mathcal{X})\alpha, \Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}_0}\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0}^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{X}}\right) \\ - \frac{1}{2}\text{trace}\left(\Lambda_{\mathcal{X}}^{-1}\left(\mathbb{K}_{\mathcal{X}\mathcal{X}} - \mathbb{K}_{\mathcal{X}\mathcal{X}_0}\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0}^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{X}}\right)\right) \end{aligned} \tag{3.7}$$

By the Woodbury matrix identity [Rasmussen & Williams (2006)], one can derive the follow-

ing:

$$\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}_0}\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0}^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{X}}\right)^{-1} = \Lambda_{\mathcal{X}}^{-1} - \Lambda_{\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}\mathcal{X}_0}\left(\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0} + \mathbb{K}_{\mathcal{X}_0\mathcal{X}}\Lambda_{\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}\mathcal{X}_0}\right)^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{X}}\Lambda_{\mathcal{X}}^{-1} \tag{3.8}$$

Because $\Lambda_{\mathcal{X}}$ is a diagonal matrix, (3.8) is efficiently computed, so as $\mathcal{L}$. Moreover, the partial derivative of $\mathcal{L}$ with respect to α is given explicitly as follows:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \varphi\left(\mathcal{X}\right)^{\mathbb{T}}\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}_0}\mathbb{K}_{\mathcal{X}_0\mathcal{X}_0}^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{X}}\right)^{-1}\left(\mathcal{Y} - \varphi\left(\mathcal{X}\right)\alpha\right) \tag{3.9}$$

Thus, one can learn α with a gradient descent using the partial derivative (3.9). Note that gradient descent on $\mathcal{L}$ for kernel hyperparameters is not a good idea: it is possible to express the partial derivatives in closed forms, but the time complexity increases so that the exact computation becomes intractable. The Markov-chain Monte Carlo (MCMC) algorithm on the Boltzmann distribution [Landau & Lifshitz (2013)] with $\mathcal{L}$ as the energy is a practical alternative.

The last issue is about the outlier classification: here the same kind of Bayesian modeling with that for benthic δ^{18}O in section 2.3.2 is applied.

Now we are prepared. Algorithm 2 summarizes the inference procedure.

Algorithm 2: GPR

```
initialize coefficient parameters and kernel hyperparameters
set all data to be inliers
while converging do
    tune kernel hyperparameters with inliers
    learn coefficient parameters with inliers
    update the variance function with inliers
    classify inliers and outliers
end
compute the posterior predictive as the regression model
return parameters, outliers and the regression model
```

3.3.2 Multimodal Gaussian Process Regression

As discussed in section B.2.3, a GPR is misspecified if the data has multiple modes over inputs. To deal with such a drawback, several works have been presented. [Tresp (2001); Rasmussen & Ghahramani (2002)] consider multiple generative processes based on the input space separation and local expert models. [Lázaro-Gredilla et al. (2012); Kaiser et al. (2018)] depend on a set of GP priors on the multiple hidden states that stand for different generative processes and more flexible probabilistic association models. [Bodin et al. (2017)] extends the input domain with continuous latent inputs to be paired with inputs and eventually marginalized out.

Inspired by the idea of [Bodin et al. (2017)], here we present a new multimodal GPR that extends the input domain with discrete latent inputs to be sampled by the Gibbs sampling (section A.4.1) or estimated by the EM algorithm with an *explicit* M-step. Let $\mathcal{L} = \{\mathrm{L}_n\}_{n=1}^{\mathrm{N}} \subseteq \{0, 1, 2, \cdots, \mathrm{C}\}$ and $\mathcal{L}_0$ be the latent and *fixed* pseudo-latent inputs that are coupled with the real and pseudo-inputs $\mathcal{X} = \{\mathrm{X}_n\}_{n=1}^{\mathrm{N}}$ and $\mathcal{X}_0$, respectively. Then, the extended kernel $\widetilde{\mathbb{K}}$ is defined as follows:

$$\widetilde{\mathbb{K}}\left(\left(x_1, l_1\right), \left(x_2, l_2\right)\right) \triangleq \mathbb{K}\left(x_1, x_2\right) \cdot \exp\left(-\partial^2 \cdot 1_{\{l_1 \neq l_2\}}\right) \tag{3.10}$$

, where ∂ is a hyperparameter that controls the cluster similarity. Note that $\widetilde{\mathbb{K}}$ is a well-defined kernel because $d\left(l_1, l_2\right) = 1_{\{l_1 \neq l_2\}}$ is a metric over $\{0, 1, 2, \cdots, \mathrm{C}\}$.

Now, the full extended model is specified by the following observational model (3.11), GP prior (3.12) and cluster prior (3.13): note that assigning a far smaller value to ω_0 than the other ω_c's implies that the cluster 0 is for outliers.

$$p\left(\mathcal{Y} | \beta, \mathcal{X}, \mathcal{L}\right) = \mathcal{N}\left(\mathcal{Y} \middle| \beta, \Lambda_{\mathcal{X}}^{(\mathcal{L})}\right) \triangleq \prod_{n=1}^{\mathrm{N}} \prod_{c=0}^{C} \mathcal{N}\left(\mathrm{Y}_n \middle| \beta_n, \Lambda_{\mathrm{X}_n}^{(c)}\right)^{1_{\{\mathrm{L}_n = c\}}} \tag{3.11}$$

$$p\left(\beta | \mathcal{X}, \mathcal{L}\right) = \mathcal{GP}\left(\beta \middle| \mu_{\mathcal{X}}^{(\mathcal{L})}, \mathbb{K}_{\mathcal{X}\mathcal{L}, \mathcal{X}\mathcal{L}}\right) \tag{3.12}$$

$$p(\mathcal{L}) = \prod_{n=1}^{\mathrm{N}} \prod_{c=0}^{\mathrm{C}} \omega_c^{1_{\{\mathrm{L}_n=c\}}} \tag{3.13}$$

Once the label $\mathcal{L}$ is given, one can compute the corresponding posterior predictive distribution similarly with (3.6). If $\mathcal{L}$ is given as a set of samples, then the associated posterior predictive is obtained as the approximation that marginalizes $\mathcal{L}$ out by averaging them over samples.

What we are interested in is how to update $\mathcal{L}$. Suppose that all the unknowns are given. One would like to directly access to the posterior of $\mathcal{L}$ as follows:

$$\begin{aligned} p(\mathcal{L}|\mathcal{X},\mathcal{Y}) \propto p(\mathcal{Y},\mathcal{L}|\mathcal{X}) &= p(\mathcal{L}) \int p(\mathcal{Y}|\beta,\mathcal{X},\mathcal{L})\, p(\beta|\mathcal{X},\mathcal{L})\, d\beta \\ &= \left(\prod_{n=1}^{\mathrm{N}} \prod_{c=0}^{\mathrm{C}} \omega_c^{1_{\{\mathrm{L}_n=c\}}} \right) \cdot \mathcal{N}\left(\mathcal{Y} \middle| \mu_{\mathcal{X}}^{(\mathcal{L})}, \mathbb{K}_{\mathcal{X}\mathcal{L},\mathcal{X}\mathcal{L}} + \Lambda_{\mathcal{X}}^{(\mathcal{L})} \right) \end{aligned} \tag{3.14}$$

The problem is that (3.14) is hard to deal with for $\mathcal{L}$ because the latter term is indecomposable with respect to $\mathcal{L}$. Instead, let us think about the following: recall that γ is assumed to be a sufficient statistic in section B.7.

$$\begin{aligned} p(\mathcal{Y}|\gamma,\mathcal{X},\mathcal{L}) &\approx \prod_{n=1}^{\mathrm{N}} \int p(\mathrm{Y}_n|b,\mathrm{X}_n,\mathrm{L}_n)\, p(b|\mathrm{X}_n,\mathrm{L}_n,\gamma,\mathcal{X}_0,\mathcal{L}_0)\, db \\ &= \prod_{n=1}^{\mathrm{N}} \mathcal{N}\left(\mathrm{Y}_n|\widetilde{\mu}(\mathrm{X}_n,\mathrm{L}_n,\gamma), \widetilde{\nu}(\mathrm{X}_n,\mathrm{L}_n)\right) \\ \widetilde{\mu}(\mathrm{X}_n,\mathrm{L}_n,\gamma) &= \mu_{\mathrm{X}_n}^{(\mathrm{L}_n)} + \mathbb{K}_{\mathrm{X}_n\mathrm{L}_n,\mathcal{X}_0\mathcal{L}_0} \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}_0\mathcal{L}_0}^{-1} \left(\gamma - \mu_{\mathcal{X}_0}^{(\mathcal{L}_0)}\right) \triangleq \widetilde{\mu}_n \\ \widetilde{\nu}(\mathrm{X}_n,\mathrm{L}_n) &= \Lambda_{\mathrm{X}_n}^{(\mathrm{L}_n)} + \mathbb{K}_{\mathrm{X}_n\mathrm{L}_n,\mathrm{X}_n\mathrm{L}_n} - \mathbb{K}_{\mathrm{X}_n\mathrm{L}_n,\mathcal{X}_0\mathcal{L}_0} \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}_0\mathcal{L}_0}^{-1} \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathrm{X}_n\mathrm{L}_n} \triangleq \widetilde{\nu}_n \end{aligned} \tag{3.15}$$

Therefore, one can approximate $p(\mathcal{L}|\gamma,\mathcal{X},\mathcal{Y})$ with following:

$$\begin{aligned} p(\mathcal{L}|\gamma,\mathcal{X},\mathcal{Y}) &\propto p(\mathcal{Y}|\gamma,\mathcal{X},\mathcal{L})\, p(\mathcal{L}) \\ &\approx \prod_{n=1}^{\mathrm{N}} \left(\mathcal{N}\left(\mathrm{Y}_n|\widetilde{\mu}(\mathrm{X}_n,\mathrm{L}_n,\gamma), \widetilde{\nu}(\mathrm{X}_n,\mathrm{L}_n)\right) \prod_{c=0}^{\mathrm{C}} \omega_c^{1_{\{\mathrm{L}_n=c\}}} \right) \end{aligned} \tag{3.16}$$

Note that (3.16) is decomposable with respect to $\mathcal{L}$, thus sampling $\mathcal{L}$ can be done efficiently

given $\mathcal{X}$, $\mathcal{Y}$ and the latent sufficient statistic γ. Together with $p\left(\gamma|\mathcal{X},\mathcal{Y},\mathcal{L}\right)\approx\varphi\left(\gamma|\mathcal{L}\right)$ in section B.7, Gibbs sampling (section A.4.1) efficiently samples γ given $\mathcal{L}$ and $\mathcal{L}$ given γ iteratively. Furthermore, one can estimate $\mathcal{L}$ by the following EM algorithm that marginalizes γ out:

E-step: Compute the following expectation:

$$\Psi\left(\mathcal{L}\middle|\mathcal{L}^{(t)}\right)\triangleq\mathbb{E}_{\gamma\middle|\mathcal{X},\mathcal{Y},\mathcal{L}^{(t)}}\left[\log p\left(\mathcal{Y}|\gamma,\mathcal{X},\mathcal{L}\right)+\log p\left(\gamma|\mathcal{X}_0,\mathcal{L}_0\right)+\log p\left(\mathcal{L}\right)\right] \tag{3.17}$$

M-step: Find $\mathcal{L}^{(t+1)}\triangleq\operatorname{argmax}_{\mathcal{L}}\Psi\left(\mathcal{L}\middle|\mathcal{L}^{(t)}\right)$.

With the approximation $p\left(\gamma|\mathcal{X},\mathcal{Y},\mathcal{L}\right)\approx\varphi\left(\gamma|\mathcal{L}\right)$, we have:

$$\begin{aligned}&\Psi\left(\mathcal{L}\middle|\mathcal{L}^{(t)}\right)=\int p\left(\gamma\middle|\mathcal{X},\mathcal{Y},\mathcal{L}^{(t)}\right)\left[\log p\left(\mathcal{Y}|\gamma,\mathcal{X},\mathcal{L}\right)+\log p\left(\gamma|\mathcal{X}_0,\mathcal{L}_0\right)+\log p\left(\mathcal{L}\right)\right]d\gamma\\&\approx\log p\left(\mathcal{L}\right)+\int\varphi\left(\gamma\middle|\mathcal{L}^{(t)}\right)\log p\left(\mathcal{Y}|\gamma,\mathcal{X},\mathcal{L}\right)d\gamma+\int\varphi\left(\gamma\middle|\mathcal{L}^{(t)}\right)\log p\left(\gamma|\mathcal{X}_0,\mathcal{L}_0\right)d\gamma\end{aligned} \tag{3.18}$$

Because the last term in (3.18) is irrelevant to $\mathcal{L}$, we just need to find the maximizer of the sum of the first and second terms. Note that $p\left(\mathcal{Y}|\gamma,\mathcal{X},\mathcal{L}\right)$ is approximated with (3.15), thus:

$$\begin{aligned}&\Psi\left(\mathcal{L}\middle|\mathcal{L}^{(t)}\right)-\int\varphi\left(\gamma\middle|\mathcal{L}^{(t)}\right)\log p\left(\gamma|\mathcal{X}_0,\mathcal{L}_0\right)d\gamma\\\approx&\log p\left(\mathcal{L}\right)\\&+\int\varphi\left(\gamma\middle|\mathcal{L}^{(t)}\right)\log\prod_{n=1}^{\mathrm{N}}\mathcal{N}\left(\mathrm{Y}_n\middle|\mu_{\mathrm{X}_n}^{(\mathrm{L}_n)}+\mathbb{K}_{\mathrm{X}_n\mathrm{L}_n,\mathcal{X}_0\mathcal{L}_0}\mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}_0\mathcal{L}_0}^{-1}\left(\gamma-\mu_{\mathcal{X}_0}^{(\mathcal{L}_0)}\right),\widetilde{\nu}_n\right)d\gamma\\=&\log\prod_{n=1}^{\mathrm{N}}\prod_{c=0}^{\mathrm{C}}\omega_c^{1_{\{\mathrm{L}_n=c\}}}\\&+\sum_{n=1}^{\mathrm{N}}\int\varphi\left(\gamma\middle|\mathcal{L}^{(t)}\right)\log\mathcal{N}\left(\mathrm{Y}_n\middle|\mu_{\mathrm{X}_n}^{(\mathrm{L}_n)}+\mathbb{K}_{\mathrm{X}_n\mathrm{L}_n,\mathcal{X}_0\mathcal{L}_0}\mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}_0\mathcal{L}_0}^{-1}\left(\gamma-\mu_{\mathcal{X}_0}^{(\mathcal{L}_0)}\right),\widetilde{\nu}_n\right)d\gamma\\=&\sum_{n=1}^{\mathrm{N}}\left[\sum_{c=0}^{\mathrm{C}}1_{\{\mathrm{L}_n=c\}}\log\omega_c-\frac{1}{2}\left(\frac{\mathrm{A}_n^2+\mathrm{B}_n}{\widetilde{\nu}_n}+\log\widetilde{\nu}_n\right)\right]\end{aligned} \tag{3.19}$$

, where:

$$
\begin{aligned}
\mathrm{A}_n &\triangleq \mathrm{Y}_n - \mu_{\mathrm{X}_n}^{(\mathrm{L}_n)} - \mathbb{K}_{\mathrm{X}_n\mathrm{L}_n,\mathcal{X}_0\mathcal{L}_0}\Sigma^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}\mathcal{L}^{(t)}}\left(\Lambda_{\mathcal{X}}^{\left(\mathcal{L}^{(t)}\right)}\right)^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}^{\left(\mathcal{L}^{(t)}\right)}\right) \\
\mathrm{B}_n &\triangleq \mathbb{K}_{\mathrm{X}_n\mathrm{L}_n^{(t)},\mathcal{X}_0\mathcal{L}_0}\Sigma^{-1}\mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathrm{X}_n\mathrm{L}_n^{(t)}} \\
\Sigma &\triangleq \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}_0\mathcal{L}_0} + \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}\mathcal{L}^{(t)}}\left(\Lambda_{\mathcal{X}}^{\left(\mathcal{L}^{(t)}\right)}\right)^{-1}\mathbb{K}_{\mathcal{X}\mathcal{L}^{(t)},\mathcal{X}_0\mathcal{L}_0}
\end{aligned}
\tag{3.20}
$$

Thus, $\Psi\left(\mathcal{L}\middle|\mathcal{L}^{(t)}\right)$ is decomposable with respect to $\mathcal{L}$. Thus, the M-step can be efficiently run by following:

$$
\mathrm{L}_n^{(t+1)} = \underset{\mathrm{L}_n}{\operatorname{argmax}}\left[\log\omega_{\mathrm{L}_n} - \frac{1}{2}\left(\frac{\mathrm{A}_n^2 + \mathrm{B}_n}{\widetilde{\nu}_n} + \log\widetilde{\nu}_n\right)\right] \tag{3.21}
$$

What remains is to deal with the unknown label l for the query input x. In the learning procedure, all labels for the inputs are learned by the EM algorithm as described above. After learning, we have a pair of inputs and the learned labels $\left(\mathcal{X}, \hat{\mathcal{L}}\right)$. Any classifier $p\left(l\middle|x; \mathcal{X}, \hat{\mathcal{L}}\right)$ including Gaussian process classifier (section B.3) that is learned from $\left(\mathcal{X}, \hat{\mathcal{L}}\right)$ can be the association distribution of the label l given the query input x.

Finally, the posterior predictive distribution is given by the following Gaussian mixture:

$$
p\left(y|x, \mathcal{X}, \mathcal{Y}, \mathcal{X}_0\right) = \sum_{l=0}^{\mathrm{C}} p\left(l\middle|x; \mathcal{X}, \hat{\mathcal{L}}\right)\mathcal{N}\left(y\middle|\mu_x^{(l)} + m\left(x, l\right), \Lambda_x^{(l)} + s\left(x, l\right)\right) \tag{3.22}
$$

, where: let $\mathbb{K}_{xl,0} = \mathbb{K}_{xl,\mathcal{X}_0\mathcal{L}_0}$, $\mathbb{K}_{0,xl} = \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,xl}$, $\mathbb{K}_{00} = \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}_0\mathcal{L}_0}$, $\mathbb{K}_{01} = \mathbb{K}_{\mathcal{X}_0\mathcal{L}_0,\mathcal{X}\hat{\mathcal{L}}}$ and $\mathbb{K}_{10} = \mathbb{K}_{\mathcal{X}\hat{\mathcal{L}},\mathcal{X}_0\mathcal{L}_0}$ for short.

$$
\begin{aligned}
m\left(x, l\right) &\triangleq \mathbb{K}_{xl,0}\left(\mathbb{K}_{00} + \mathbb{K}_{01}\left(\Lambda_{\mathcal{X}}^{\left(\hat{\mathcal{L}}\right)}\right)^{-1}\mathbb{K}_{10}\right)^{-1}\mathbb{K}_{01}\left(\Lambda_{\mathcal{X}}^{\left(\hat{\mathcal{L}}\right)}\right)^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}^{\left(\hat{\mathcal{L}}\right)}\right) \\
s\left(x, l\right) &\triangleq \mathbb{K}_{xl,xl} - \mathbb{K}_{xl,0}\mathbb{K}_{00}^{-1}\mathbb{K}_{0,xl} + \mathbb{K}_{xl,0}\left(\mathbb{K}_{00} + \mathbb{K}_{01}\left(\Lambda_{\mathcal{X}}^{\left(\hat{\mathcal{L}}\right)}\right)^{-1}\mathbb{K}_{10}\right)^{-1}\mathbb{K}_{0,xl}
\end{aligned}
\tag{3.23}
$$

Now we are again prepared. Algorithm 3 summarizes the inference procedure: note that this algorithm can be reduced to an unsupervised classifier by discarding last two steps before the last

line and not returning the association distribution and regression model (removable steps).

Algorithm 3: Multimodal GPR

```
initialize coefficient parameters and kernel hyperparameters
randomize the labels
while converging do
    update labels by the EM algorithm
    tune kernel hyperparameters
    for each label do
        learn the label-specific mean prior function with the classified data if needed
        update the label-specific variance function with the classified data
    end
end
learn association distribution given the learned labels (removable)
compute the posterior predictive as the regression model (removable)
return parameters, labels, association distribution (removable) and the regression
  model (removable)
```

Figure 3.2 visualizes the results obtained by running the reduced version of algorithm 2 on some examples. We could check that similar result was obtained with that of [Kaiser et al. (2018)] to classify outliers in the last panel of the figure. Though [Bodin et al. (2017)] deals with a more general setting that the extended inputs are continuous and to be marginalized out, our model so far is not only mathematically but also practically interesting because it efficiently updates the labels by the EM algorithm with an explicit M-step or samples them by the Gibbs sampling algorithm.

3.4 Results and Discussion

For each proxy, we tried the following three models: homoscedastic GPR, heteroscedastic GPR and the reduced multimodal GPR for finding multiple clusters of the data if exist. Each set of results are visualized in the corresponding figure that consists of six panels. The first panel shows the resulting calibration model of the proxy over SST; the second panel presents the nonparametric bias function over SST; the third panel indicates the locations of inliers (green) and outliers (red) on the world map; the fourth panel gives the inferred standard deviation function over SST;

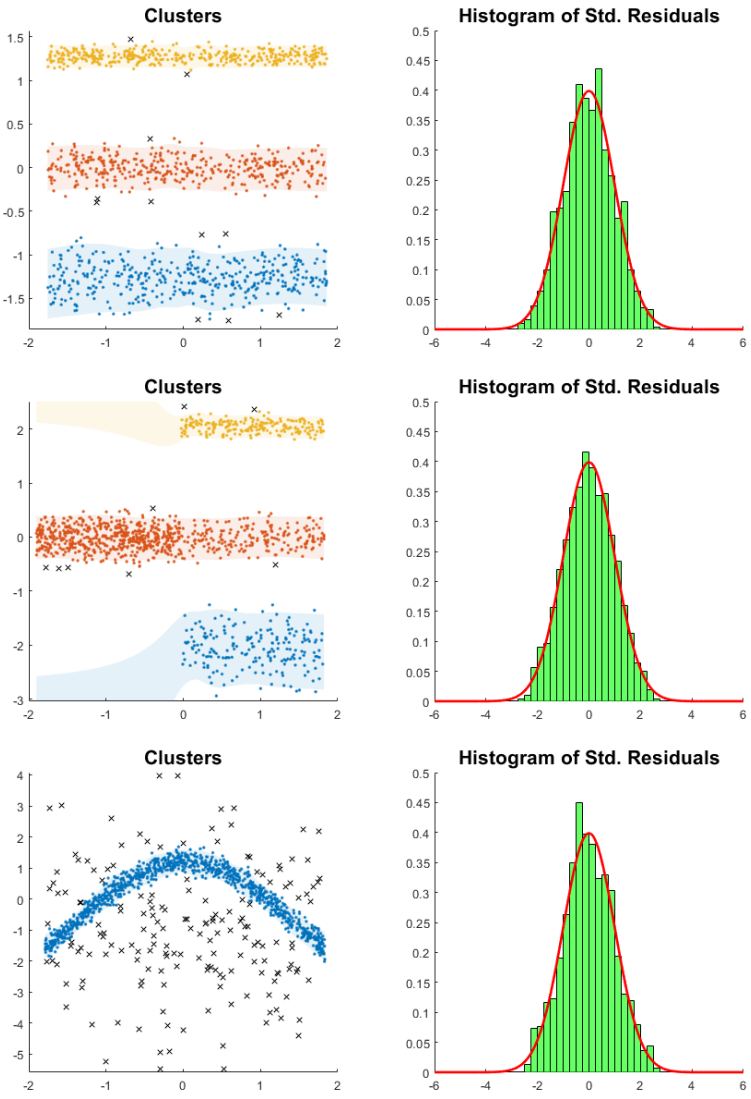


Figure 3.2: Some example results from our multimodal GPR.

the fifth and sixth panels are about the standardized residuals that are obtained by subtracting the proxy observations by their regressions and then dividing them by their inferred standard deviations.

3.4.1 Alkenone Calibration Models

Figure 3.3 visualizes the homoscedastic GPR results for the alkenone proxy data. As shown in the first and second panels, too many observations are classified as outliers, which implies that the model is not correctly specified. It is also supported by the histogram of the standardized residuals: inferred standard deviation function seems to be too narrow to encompass all the observations, but also too broad for the classified inliers. The classified outliers show somehow but not certain regional pattern in the world map. Even though the model was homoscedastic GPR, the inferred standard deviation function over SST is heteroscedastic: because observational variance is constant, this inconsistency stems from the model uncertainty. The unstable nonparametric bias function also supports the existence of extensive model uncertainty.

The heteroscedastic GPR results are visualized in figure 3.4. In contrast to the homoscedastic case, the nonparametric bias function is stable and not many outliers are classified. The inferred standard deviation function over SST clearly shows the heteroscedastic characteristic of the proxy data: $U_{37}^{K'}$ observations at higher temperatures are more certain than those at lower temperatures. The histogram that fits well for the standard normal distribution also supports that the classified inliers follow a single Gaussian model. However, there is one doubtful aspect that leads to the further investigation: outliers under the regression function seem to form another cluster, though they do not show a vivid regional pattern.

Some multimodal GRP clustering results are shown in figure 3.5: those results are preliminary because the label uncertainty was not considered. Here the cluster-specific mean prior functions are defined by different constant functions. The outlier cluster mentioned in the above paragraph gets absorbed into inliers as the number of clusters increases. Note that the analysis for 1 cluster (the top panel) suggests a nonparametric way of classifying outliers.

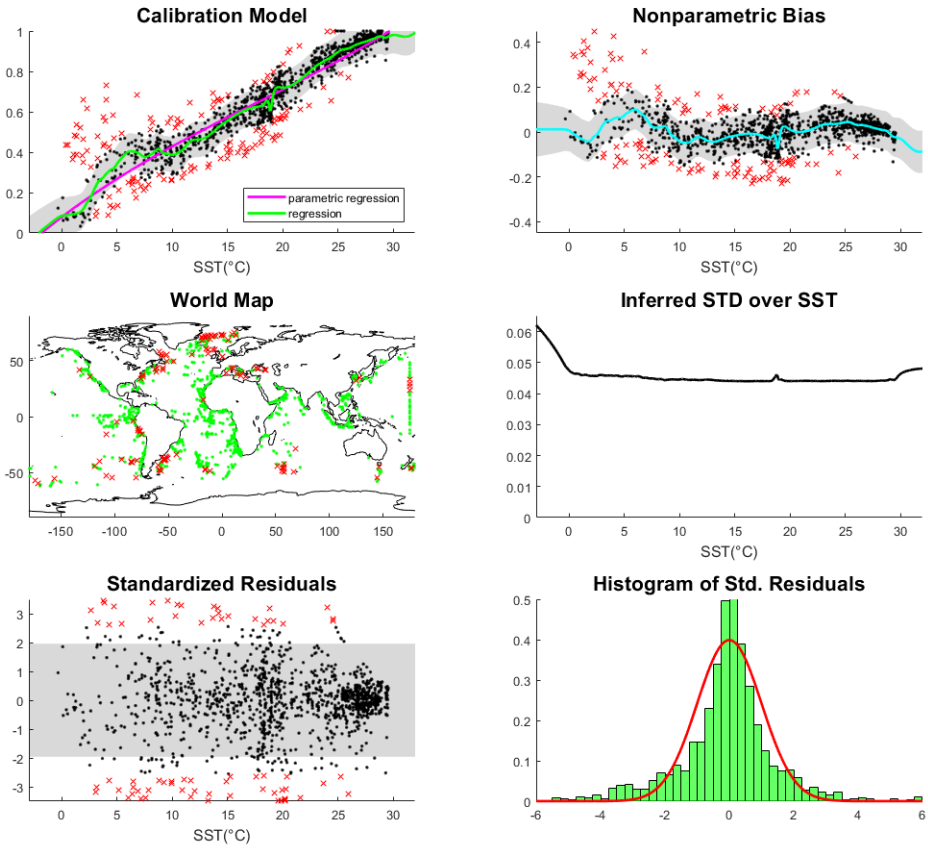


Figure 3.3: Homoscedastic Gaussian process regression of $U_{37}^{K'}$ on SST. In each panel, red crosses are the outliers and shaded region indicates the 95% confidence band.

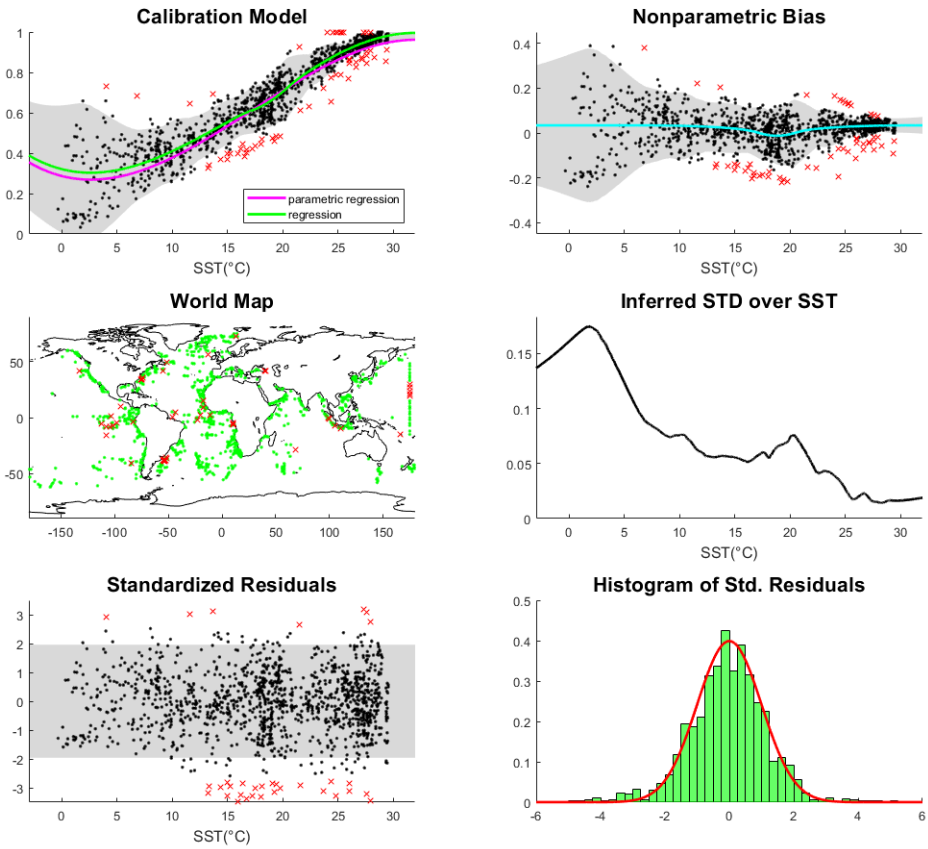


Figure 3.4: Heteroscedastic Gaussian process regression of $U_{37}^{K'}$ on SST.

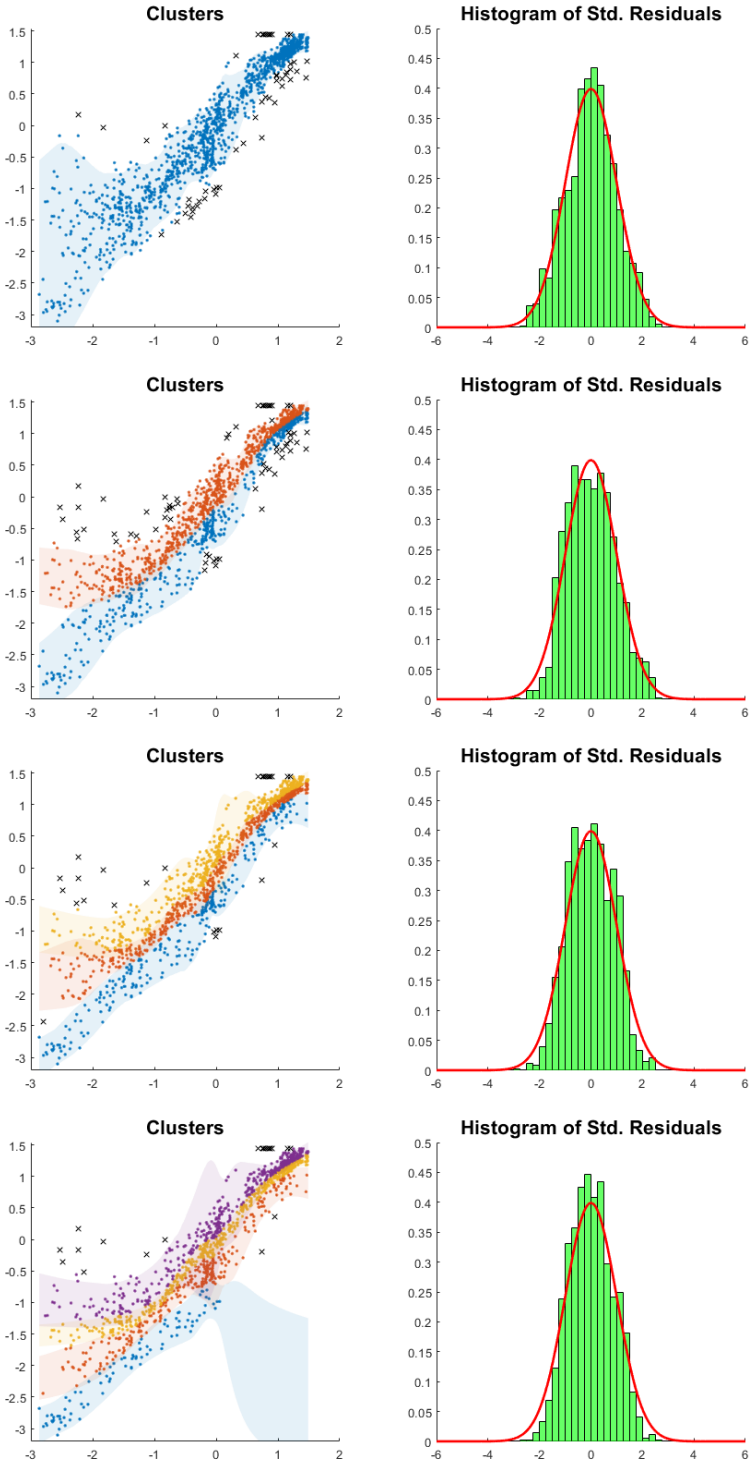


Figure 3.5: Multimodal Gaussian process regressions of $U_{37}^{K'}$ on SST.

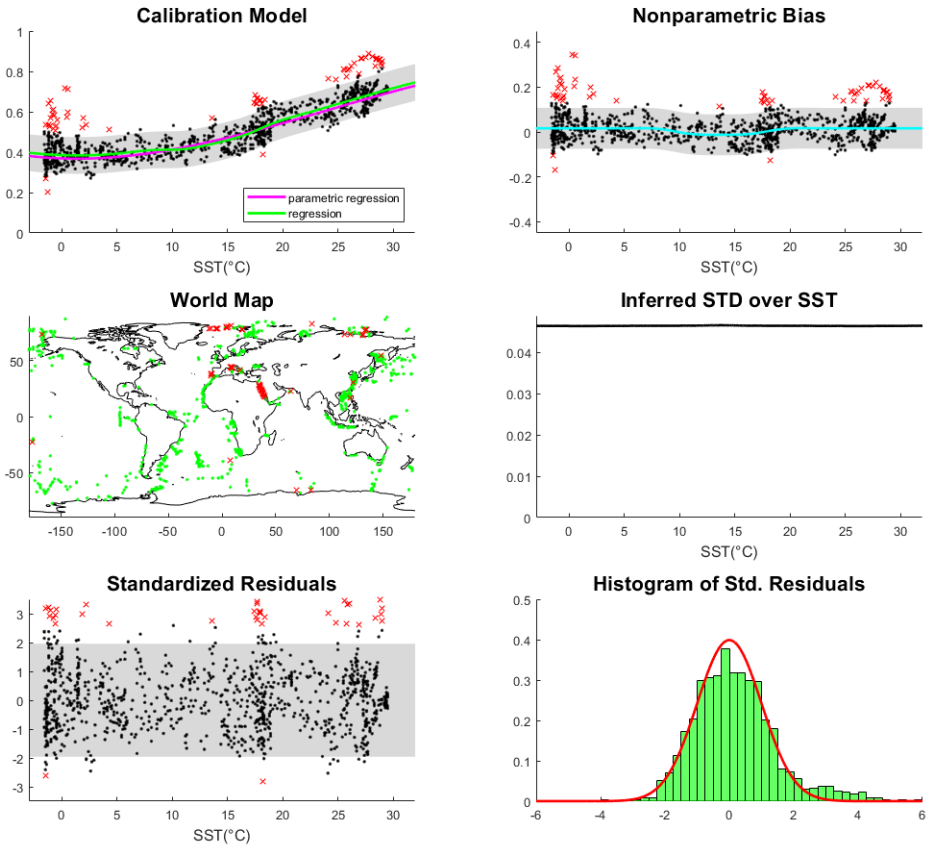


Figure 3.6: Homoscedastic Gaussian process regression of TEX_{86} on SST.

3.4.2 TEX_{86} Calibration Models

Figure 3.6 visualizes the homoscedastic GPR results for the TEX_{86} proxy data. In the world map, the outliers show a vivid regional pattern: they are mainly from the Red Sea and the Arctic Ocean. It supports that the data are commonly follow the same calibration model. Unlike the homoscedastic GPR of the alkenone in section 3.4.1, the nonparametric bias function is stable over SST, so here the regression and the parametric regression of a cubic polynomial are almost identical. However, this calibration model is still not correctly specified: the shape of the histogram of standardized residuals implies that the inliers are more likely following a Gaussian mixture rather than a single Gaussian.

The heteroscedastic GPR results are visualized in figure 3.7. The calibration model gives a

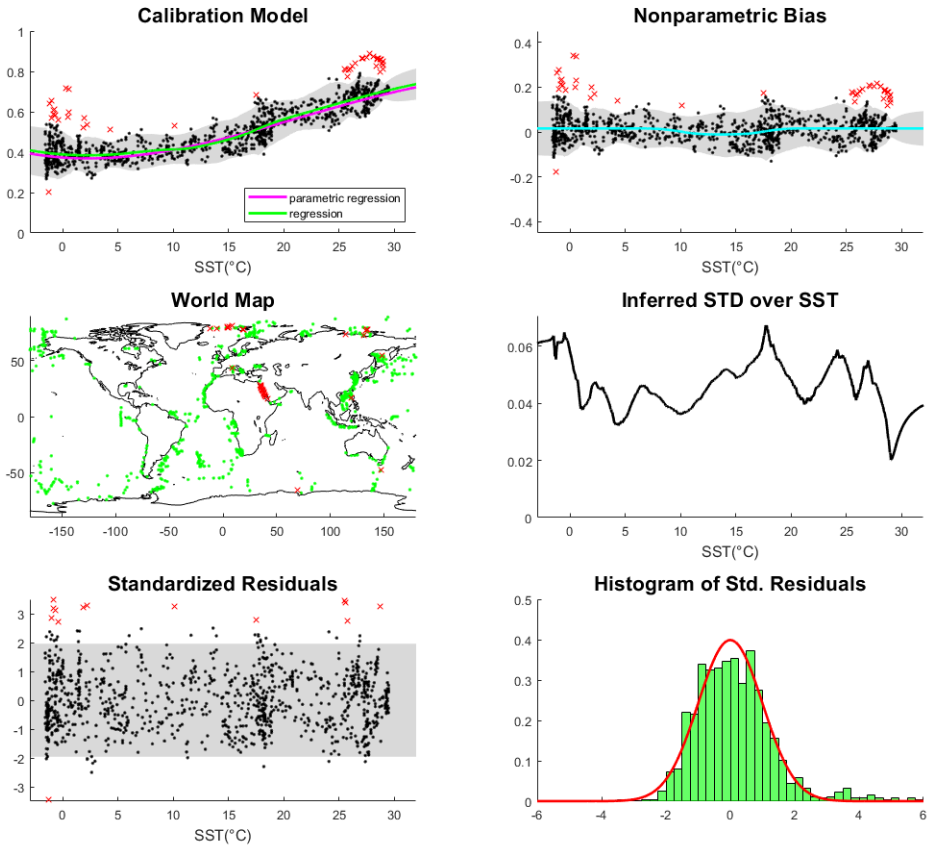


Figure 3.7: Heteroscedastic Gaussian process regression of TEX_{86} on SST.

more flexible variance function over SST to classify less observations as outliers. Thus, the regional pattern of the outliers become more vivid than the homoscedastic case. Unlike alkenone that show the diminishing pattern, the inferred standard deviation function of TEX_{86} over SST does not have a certain pattern here. Though outliers are addressed more carefully and sensitively, the histogram of standardized residuals still implies that the observations tend to follow a mixture of Gaussian distributions.

Some preliminary multimodal GRP clustering results are shown in figure 3.8. Histograms of the standardized residuals tend to fit well for the standard normal distribution as the number of clusters increases.

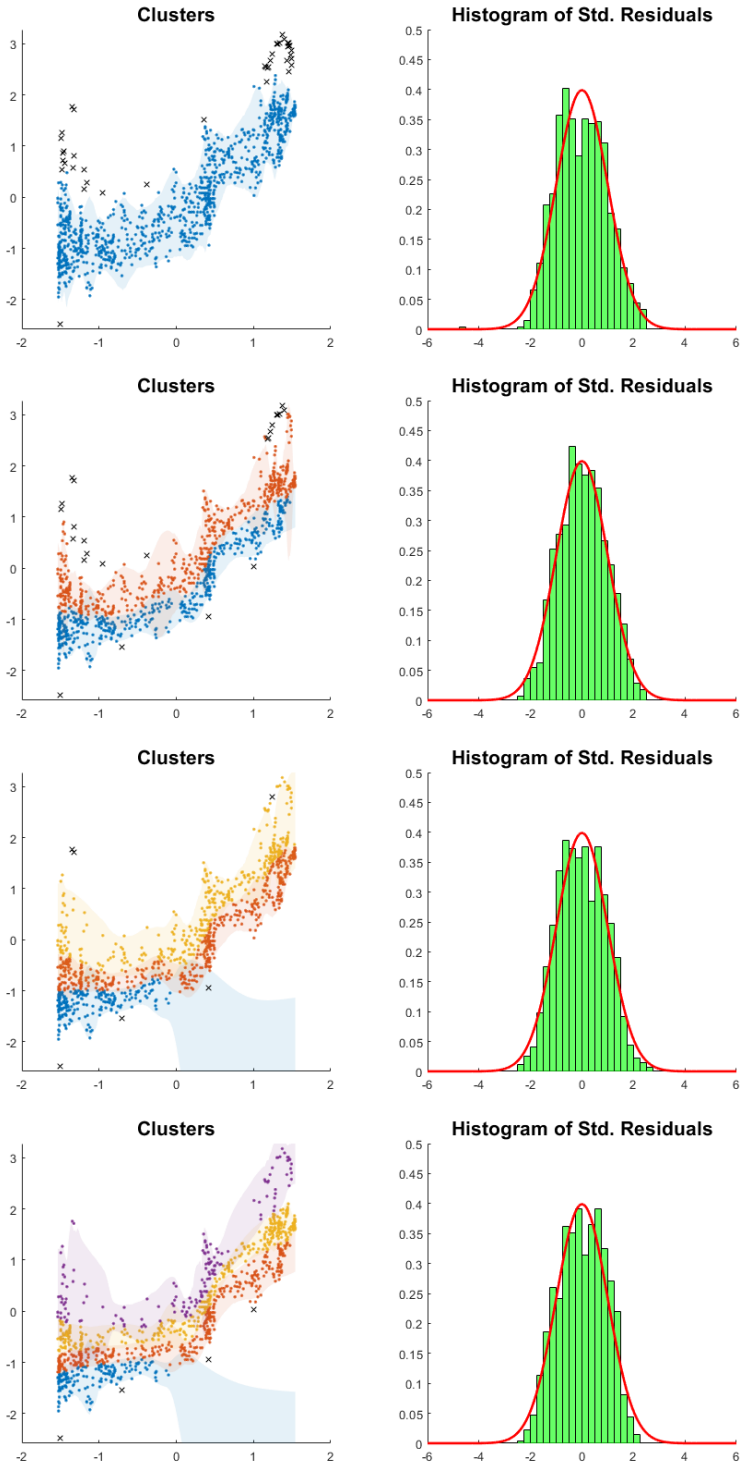


Figure 3.8: Multimodal Gaussian process regressions of TEX_{86} on SST.

3.5 Future Work

Many previous works of the calibration model construction have discarded or adjusted the proxy observations or SSTs based on the data-specific characteristics. This can be understood in the following framework: there are two categories of outliers. Outliers of the first category are just meaningless noises without any information. Thus, such outliers should be discarded in learning procedure, especially for the nonparametric models that are directly depending on the data. Outliers of the second category are, however, originally informative but structurally contaminated for some data-specific reasons. Classification into these two categories cannot be done without the aid of relevant experts in paleoceanography and/or paleoclimatology. Thus, to reflect such subtle differences of outliers, in future our work must be collaborated with the relevant experts to construct not only mathematically but also practically meaningful calibration models.

In relation to the previous criticism, an interdisciplinary work to deal with the lab-specific biases in alkenone is ongoing. One of the collaborators, Dr. Timothy Herbert, suspects that lab-specific biases exist and affect the pattern of data for the reason that the proxy dataset is a collection of the observations from multiple laboratories and each lab has its own procedure. This problem is similar to the multimodal GPR in section 3.3.2 but could be easier because it is supervised: lab associations are given a priori to be paired with the data.

As BAYSPAR of [Tierney & Tingley (2014)], including spatial information of the data to the inputs brings another interesting goal. GP models are easily extensible to take more inputs, such as spherical coordinates in this problem. In fact, we already have some preliminary results that are shown in figure 3.9: whereas BAYSPAR gives GP priors on the slope and intercept coefficients that are defined regionally with grids, our model is more nonparametric and returns the exact posterior predictive distributions instead of the slopes and intercepts. The site-specific regression model at each site tends to follow the observations near to the site over the SSTs that are in range of where those observations are made, but soon gets deviated to the general data as going outside of the range. One major problem of this approach is regularization by the guidance of experts

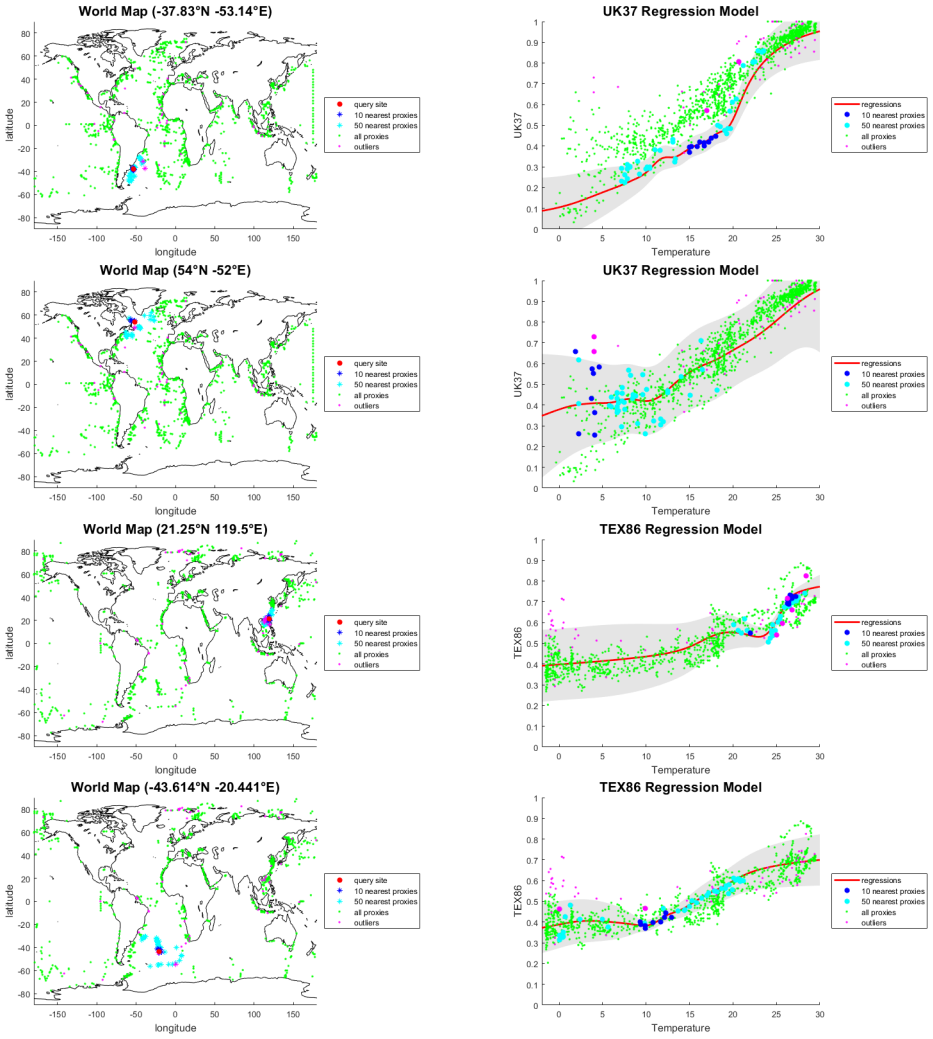


Figure 3.9: Some preliminary results from the site-specific GPR calibration model.

relevant to data: models that are too faithful to the data could lose something important such as its applicability or practicability by overfitting.

More abstract GP-based models, such as deep GP (section B.6), could be applied to the datasets as one of the future research directions. Like what deep neural network models do, deep GPs are simultaneously doing inference and extracting features. As discussed in section B.6, deep GPs contribute to overcoming the suboptimality of GP models, at the cost of interpretability: shallow GPR models are given explicitly and mean and variance terms of the posterior predictive distributions can be decomposed into either parametric regression and nonparametric bias or observational uncertainty and model uncertainty, as discussed earlier, but a deep GP regression

cannot be expressed in a closed form thus loses such a decomposable aspect.

This project ends at the calibration model construction: the obvious work to be followed is constructing a model for reconstructing SSTs given ancient proxy data. It could be done independently on the proxies, i.e., restoring an SST for each proxy one-by-one, but a better approach would be connecting SSTs over ages or locations, as discussed in section 3.2. If we define a GP state-space model over ages and locations, then the choice of kernels or their hyperparameters could be suggested by the relevant experts, based on their physical models from other related data.

4

Atmospheric CO_2 Reconstruction Using Boron Proxy

Carbon dioxide (CO_2) is a colorless gas in Earth's atmosphere, consists of a carbon atom covalently double bonded to two oxygen atoms. The current concentration is about 0.04% (412 ppm) by volume, having risen from pre-industrial levels of 280 ppm [Eggleton (2012)]. Though CO_2 gas occupies only a small portion of the atmosphere, its variations are key factors of climate change that has recently been a serious issue threatening all mankind.

The boron isotope, which is also a proxy for pH reconstruction, is one of the main marine proxies used for the atmospheric CO_2 reconstruction. Boron has two stable isotopes, ^{11}B and

^{10}B, in the forms of boric acid $B(OH)_3$ and borate ion $B(OH)_4^-$. Because boric acid has a higher $^{11}B/^{10}B$ ratio than borate ion by 1.0272 [Klochko et al. (2006)], the boron isotopic composition in marine carbonate is affected by the pH of the seawater in which the carbonate grows: the boron proxy is based on the observation that predominantly only borate ion is incorporated into marine calcium carbonate, such as the shells of foraminifera [Hemming & Hanson (1992)].

The composition of total dissolved carbon (DIC) to total alkalinity (ALK) is linked to the seawater pH that is closely related to CO_2. Seawater CO_2 content is estimated by the nature of the CO_2-system (with additional parameters) in seawater that considers the boron isotope composition [Lemarchand et al. (2002)] and minor vital effects relating to well-preserved foraminiferal physiology [Henehan et al. (2016)]. Atmospheric CO_2 is then reconstructed from the seawater CO_2 by taking the air-sea disequilibrium that should be accounted for as well*.

This chapter covers our preliminary work on the reconstruction of atmospheric CO_2 with boron proxy. In contrast to the previous work that accesses each proxy observation one-by-one, we apply various state-space models such as the particle smoother (section A.3.2) and Gaussian process state-space (GPST) model (section B.4).

4.1 Related Work

While [Hönisch et al. (2009)] estimated a 2.1-million-year-long sea surface partial pressure of CO_2 (pCO_2) from boron isotopes in planktic foraminifer shells and could confirm that their estimates are consistent with the ice core records, [Foster & Rae (2016)] more carefully described how CO_2 is reconstructed from the boron isotopic composition ($\delta^{11}B$) of marine calcium carbonate. They reviewed the chemical principles underlying the proxy, summarized the available calibration data, and gave details of how boron isotopes can be utilized for estimating ocean pH and ultimately past atmospheric CO_2.

While the above works have been focusing on the pointwise inference, [Garcia-Olivares & Herrero (2013)] constructed Local Stratification (LS) model, a dynamic model of global ice volume,

*I refer to http://www.p-co2.org/boron for the first three paragraphs in this chapter.

extent of the Antarctic ice sheet and atmospheric CO_2 over time. The LS model connects a set of different past events by the system of first-order ordinary differential equations (ODEs), just as PP04 models of [Paillard & Parrenin (2004); Garcia-Olivares & Herrero (2012)] that were fitted with the experimental time series of $\delta^{18}O$ and CO_2 for the last 800 kiloyears.

4.2 Research Motive and Data

As discussed at the beginning of chapter A, the inference on the linked hidden states (past climate events over time) given the paired observations (proxy observations) can be effectively processed in the framework of state-space models. Thus, our goal is to design and run various state-space models for inferring the atmospheric CO_2 using boron proxy and verifying the results with the Antarctic ice core records [Monnin et al. (2001); Petit et al. (1999); Raynaud et al. (2005); Indermühle et al. (2000); Siegenthaler et al. (2005); Lüthi et al. (2008)] that have trapped air composed with CO_2 thus to be used as the "true" values. Because such "true" values are limited up to 800 kiloyears, our inference is also done up to at most that point. Let us call them "experimental time series".

A transition model of the state-space model can be either parametric or nonparametric. In this section we consider both of the cases. Firstly, we construct a parametric state-space model with the transition model based on the dynamic model on multiple climate events including atmospheric CO_2 of the LS model [Garcia-Olivares & Herrero (2013)]: its parametric form is justified scientifically thus itself meaningful. This work was collaborated with my advisor Dr. Charles E. Lawrence and one of his students, Andrew Kaluzny and presented to the attendees of the 2nd Workshop on Cenozoic pCO_2 Reconstruction by my advisor on May 30th, 2018.

However, the state-space model with a parametric transition model would be problematic for the following reasons: 1) the inference highly depends on the estimated parameters of the transition model, but the underlying dynamics could be nonstationary due to the change of the other surrounding conditions over time, and 2) the transition of the hidden states is usually assumed to be made regularly (for example, for every kiloyear), but the boron proxy is not regularly spaced

over ages and too scarce to be properly rearranged regularly. The GPST is a nonparametric state-space model that deals with such irregularly-spaced inputs, thus could be perfect for this purpose. Therefore, we also try the GPST on the dataset.

However, the state-space model with a parametric transition model would be problematic for the following reasons: 1) the inference highly depends on the estimated parameters of the transition model, but the underlying dynamics could be nonstationary due to the change of the other surrounding conditions over time, and 2) the transition of the hidden states is usually assumed to be made regularly (for example, for every kiloyear), but the boron proxy is not regularly spaced over ages and too scarce to be properly rearranged regularly. The GPST is a nonparametric state-space model that deals with such irregularly-spaced inputs, thus could be perfect for this purpose. Therefore, we also try the GPST on the dataset.

We use the boron proxy data in ODP668 and ODP999 [Dyez et al. (2018)]. Our sea level data are referred to [Spratt & Lisiecki (2016)]. Benthic δ^{18}O proxy are from the ocean sediment cores MD95-2042 [Bard et al. (2004); Shackleton et al. (2000, 2004)], MD02-2575 [Nürnberg et al. (2008)] and MD02-2589 [Molyneux et al. (2007); Diz et al. (2007)].

Our work here is, however, strictly limited for the following reasons: 1) our emission (observation) models for proxies are just simple and fixed parametric models that are pre-trained based on the samples from other datasets, 2) some uncertain values such as ages of the proxies are regarded as the fixed ones according to the datasets, and 3) the number of boron proxy observations up to 800 kiloyears is only 86. Thus, the work is preliminary rather than complete, just for drawing the interdisciplinary experts' interests in future research opportunities.

4.3 Models

Two models are applied separately. The first model uses the particle smoother and the second one is GPST. In each model, its emission model is defined by a set of fixed observational models that are pre-trained by the data. However, the parametric transition model of the first model based is trained during the inference, and so as the kernel hyperparameters in the second model.

The first model is depending on the proxies of benthic δ^{18}O and the boron proxy, δ^{11}B, which is a variation derived from the boron isotopes as follows [Foster & Rae (2016)]:

$$\delta^{11}\mathrm{B} \triangleq \left(\left(\frac{{}^{11}\mathrm{B}/{}^{10}\mathrm{B}_{\mathrm{sample}}}{{}^{11}\mathrm{B}/{}^{10}\mathrm{B}_{\mathrm{standard}}} \right) - 1 \right) \times 1000 \tag{4.1}$$

, where ${}^{11}\mathrm{B}/{}^{10}\mathrm{B}_{\mathrm{standard}} = 4.04367$ is the boron isotopic composition of National Institute of Standards and Technology Standard Reference Material 951 boric acid [Catanzaro et al. (1970)]. We assume that any proxy observation is paired with a given age. Unevenly gathered observations are adjusted to the grid of every 1 kiloyear. Here we limit the inference up to 260 kiloyears ago, where both benthic δ^{18}O and δ^{11}B observations in the cores exist.

Whereas the first model assumes that the ice volume is another hidden state to infer, it is given a priori as the values derived from the sea level data in [Spratt & Lisiecki (2016)] and the following relationship[†]:

$$\Delta \mathrm{V} \cdot d_{\mathrm{ice}} \equiv -\Delta h \cdot \mathrm{A} \cdot d_{\mathrm{seawater}} \tag{4.2}$$

, where V is the ice volume, $\mathrm{A} = 3.618 \times 10^8 (km^2)$ is how much area the oceans cover, h is the height, and densities of ice and sea water are $d_{\mathrm{ice}} = 0.9167 (Gt/km^3)$ and $d_{\mathrm{seawater}} = 1.027 (Gt/km^3)$, respectively.

The same emission model of δ^{11}B given atmospheric CO_2 is used, whereas that of ice volume given atmospheric CO_2 is pre-trained by the paired values of the given ice volumes and given CO_2 of the Antarctic ice core records: this aspect makes the inference of this model preliminary. Here we try four kernels: squared-exponential (SE), Ornstein-Uhlenbeck (OU) and two other Matérn kernels with $d = 3/2$ and $5/2$ (section B.1.1) that correspond to the continuous autoregressive (AR) models AR (∞), AR (1), AR (2) and AR (3), respectively. The inference is done up to 800 kiloyears ago.

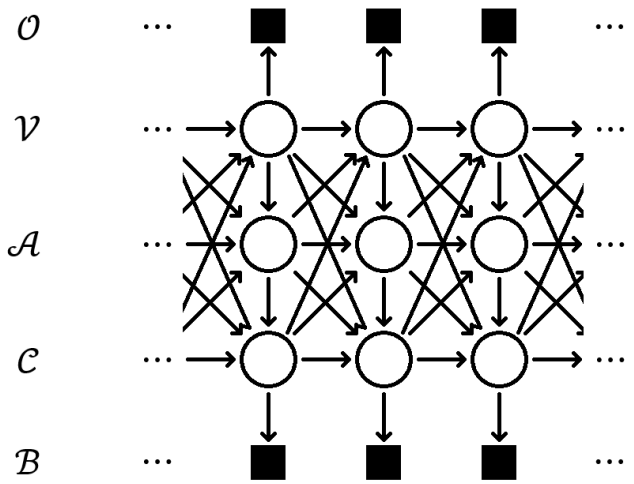


Figure 4.1: A graphical representation of our particle smoother for the CO_2 reconstruction.

4.3.1 Particle Smoother

First of all, we define the notations as follows: a graphical representation of our particle smoother is in figure 4.1.

- $\mathcal{V} = \{V_t\}_{t=1}^{T}$: dimensionless indices of ice volumes.
- $\mathcal{A} = \{A_t\}_{t=1}^{T}$: dimensionless indices of extents of the Antarctic ice sheet.
- $\mathcal{C} = \{C_t\}_{t=1}^{T}$: dimensionless indices of atmospheric CO_2.
- $\mathcal{S} = \{S_t\}_{t=1}^{T}$: tuples of above dimensionless indices, where each $S_t \triangleq (V_t, A_t, C_t)^{\mathbb{T}}$.
- $\mathcal{O} = \left\{O^{(m)}\right\}_{m=1}^{M_{\mathcal{O}}}$: set of observed $\partial^{18}O$ proxies, where each $O^{(m)} = \left\{O_t^{(m)}\right\}_{t=1}^{T}$ from core m is the set of either interpolated values to be paired with $\mathcal{S}$ or $\emptyset$ (we do not extrapolate the data).
- $\mathcal{B} = \left\{B^{(m)}\right\}_{m=1}^{M_{\mathcal{B}}}$: set of observed $\partial^{11}B$ proxies, where each $B^{(m)} = \left\{B_t^{(m)}\right\}_{t=1}^{T}$ from core m is the set of either interpolated values to be paired with $\mathcal{S}$ or $\emptyset$ (we do not extrapolate the data).

[†]http://www.antarcticglaciers.org/glaciers-and-climate/estimating-glacier-contribution-to-sea-level-rise/

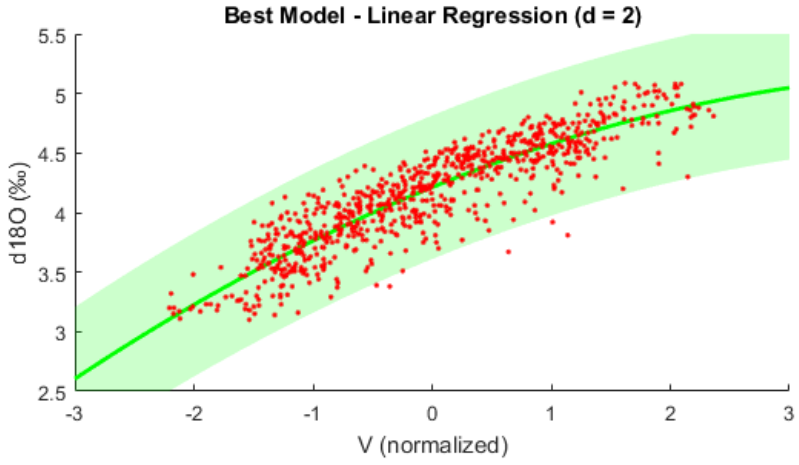


Figure 4.2: Ice volume and δ^{18}O pairs (red) and the chosen model with the 95% confidence band (green).

From the datasets, we first learn the following parametric observational model of δ^{18}O (O) given the dimensionless indices of ice volume (V):

$$p_{\mathcal{E}_V}(\mathrm{O}|\mathrm{V}) = \mathcal{N}\left(\mathrm{O}\middle|\beta_0 + \beta_1\mathrm{V} + \cdots + \beta_d\mathrm{V}^d, \exp\left(\gamma_0 + \gamma_1\mathrm{V} + \cdots + \gamma_c\mathrm{V}^c\right)\right) \tag{4.3}$$

Based on the Bayesian information criterion (BIC) [Schwarz (1978)], we choose the degree $d = 2$ and $c = 0$. The pre-trained emission model is given as follows: see figure 4.2 for its visualization with training data.

$$p_{\mathcal{E}_V}(\mathrm{O}|\mathrm{V}) = \mathcal{N}\left(\mathrm{O}\middle|4.21 + 0.407\mathrm{V} - 0.0429\mathrm{V}^2, 0.197^2\right) \tag{4.4}$$

Also, we choose the observational model of δ^{11}B (B) given the dimensionless indices of atmospheric CO_2 (C) among the following two parametric models: these models with training samples were suggested and gathered by Andrew Kaluzny.

$$\begin{aligned} p_{\mathcal{E}_B}(\mathrm{B}|\mathrm{C}) &= \mathcal{N}\left(\mathrm{B}\middle|a - \sqrt{b + c\log(\mathrm{C}+d)}, \exp\left(\delta_0 + \delta_1\mathrm{C} + \cdots + \delta_q\mathrm{C}^q\right)\right) \\ p_{\mathcal{E}_B}(\mathrm{B}|\mathrm{C}) &= \mathcal{N}\left(\mathrm{B}\middle|a + b\mathrm{C} + c\log(\mathrm{C}+d), \exp\left(\delta_0 + \delta_1\mathrm{C} + \cdots + \delta_q\mathrm{C}^q\right)\right) \end{aligned} \tag{4.5}$$

By applying the BIC again, we choose the following model: see figure 4.3 for its visualization

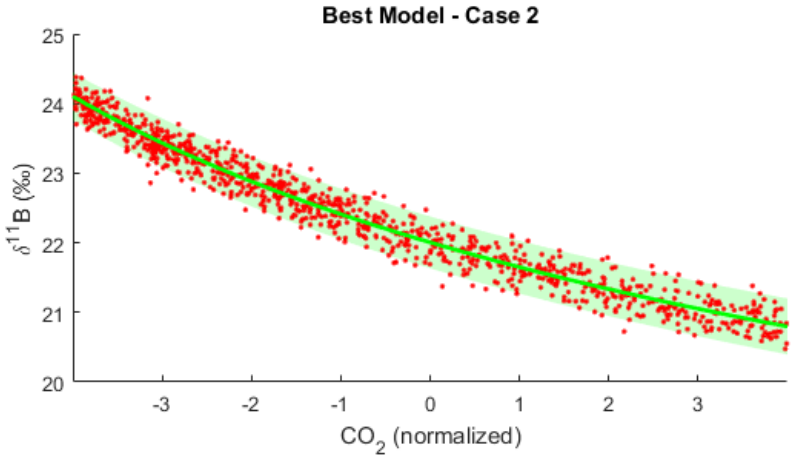


Figure 4.3: Atmospheric CO_2 and $\delta^{11}B$ pairs (red) and the chosen model with the 95% confidence band (green).

with training data.

$$p_{\mathcal{E}_\mathcal{B}}(\mathrm{B}|\mathrm{C}) = \mathcal{N}(\mathrm{B}|30.3 + 0.0552\mathrm{C} - 3.80\log(\mathrm{C} + 8.76), \exp(-3.32 + 0.0392\mathrm{C})) \quad (4.6)$$

To apply the observational models (4.4) and (4.6) for the proxy observations, just as what we did in section 2.3.2, core-specific adjustments are necessary: we linearly adjust the models by core-specific coefficient parameters.

The emission model $p_e(\mathrm{O}_t, \mathrm{B}_t|\mathrm{S}_t)$ is obtained as follows:

$$\begin{aligned} p_e\left(\mathrm{O}_t^{(m)}, \mathrm{B}_t^{(m)}\middle|\mathrm{S}_t\right) &\triangleq p_{\mathcal{E}_\mathcal{V}}\left(\mathrm{O}_t^{(m)}\middle|\mathrm{V}_t;\theta^{(m)}\right)^{1\left\{\mathrm{O}_t^{(m)}\neq\emptyset\right\}} p_{\mathcal{E}_\mathcal{B}}\left(\mathrm{B}_t^{(m)}\middle|\mathrm{C}_t;\theta^{(m)}\right)^{1\left\{\mathrm{B}_t^{(m)}\neq\emptyset\right\}} \\ p_e(\mathrm{O}_t, \mathrm{B}_t|\mathrm{S}_t) &= \prod_{m=1}^{\mathrm{M}} p_e\left(\mathrm{O}_t^{(m)}, \mathrm{B}_t^{(m)}\middle|\mathrm{S}_t\right) \end{aligned} \quad (4.7)$$

The transition model is originated by the LS model in [Garcia-Olivares & Herrero (2013)] that is defined as the system of the following first-order ODEs: $\tau_\mathrm{V}^{(1)}$, $\tau_\mathrm{V}^{(2)}$, τ_A, $\tau_\mathrm{C}^{(1)}$ and $\tau_\mathrm{C}^{(2)}$ are

parameters.

$$\begin{aligned}\frac{d\mathrm{V}}{dt} &= \frac{\mathrm{V}_r - \mathrm{V}}{\tau_{\mathrm{V}}^{(1)} 1_{\{\mathrm{V}<\mathrm{V}_r\}} + \tau_{\mathrm{V}}^{(2)} 1_{\{\mathrm{V}\geq\mathrm{V}_r\}}} \\ \frac{d\mathrm{A}}{dt} &= \frac{-\mathrm{C} - \mathrm{A}}{\tau_{\mathrm{A}}} \\ \frac{d\mathrm{C}}{dt} &= \frac{\mathrm{C}_r - \mathrm{C}}{\tau_{\mathrm{C}}^{(1)} 1_{\{\mathrm{C}<\mathrm{C}_r\}} + \tau_{\mathrm{C}}^{(2)} 1_{\{\mathrm{C}\geq\mathrm{C}_r\}}}\end{aligned} \tag{4.8}$$

, where, for the mean daily insolation I_{65} at 65°N on June 21st and parameters $x, y, z, \alpha, \beta, \gamma$ and δ,

$$\begin{aligned}\mathrm{V}_r &\triangleq -x\mathrm{C} - y\mathrm{I}_{65} + z \\ \mathrm{C}_r &\triangleq \alpha\mathrm{I}_{65} - \beta\mathrm{V} + \gamma 1_{\{b\mathrm{A}+c\mathrm{C}-d>0\}} + \delta\end{aligned} \tag{4.9}$$

By $d\mathrm{X}/dt \approx (\mathrm{X}_{n+1} - \mathrm{X}_n)/(t_{n+1} - t_n)$, one can convert (4.8) to the probabilistic transition model as follows:

$$p_t\left(\mathrm{S}_{t+1} | \mathrm{S}_t, \zeta_t\right) \triangleq \mathcal{N}\left(\mathrm{S}_{t+1} \middle| \mathrm{W}_{\zeta_t}\mathrm{S}_t + \mathrm{B}_{\zeta_t}, \Sigma\right) \tag{4.10}$$

, where Σ is a covariance matrix addressing errors and:

$$\mathrm{W}_{\zeta_t} \triangleq \begin{pmatrix} 1 - \frac{\varepsilon}{\tau_{\mathrm{V}}^{(\zeta_t)}} & 0 & -\frac{x\varepsilon}{\tau_{\mathrm{V}}^{(\zeta_t)}} \\ 0 & 1 - \frac{\varepsilon}{\tau_{\mathrm{A}}} & -\frac{\varepsilon}{\tau_{\mathrm{A}}} \\ -\frac{\beta\varepsilon}{\tau_{\mathrm{C}}^{(\zeta_t)}} & 0 & 1 - \frac{\varepsilon}{\tau_{\mathrm{C}}^{(\zeta_t)}} \end{pmatrix}, \quad \mathrm{B}_{\zeta_t} \triangleq \begin{pmatrix} -\frac{y\varepsilon}{\tau_{\mathrm{V}}^{(\zeta_t)}}\mathrm{I}_t^{65} + \frac{z\varepsilon}{\tau_{\mathrm{V}}^{(\zeta_t)}} \\ 0 \\ \frac{\alpha\varepsilon}{\tau_{\mathrm{C}}^{(\zeta_t)}}\mathrm{I}_t^{65} + \frac{\delta\varepsilon}{\tau_{\mathrm{C}}^{(\zeta_t)}} + \frac{\gamma\varepsilon}{\tau_{\mathrm{C}}^{(\zeta_t)}} \cdot 1_{\{b\mathrm{A}_t+c\mathrm{C}_t-d>0\}} \end{pmatrix}$$

$$\zeta_t \triangleq \begin{cases} 1, & \mathrm{V}_t < -x\mathrm{C}_t - y\mathrm{I}_t^{65} + z \quad \mathrm{C}_t < \alpha\mathrm{I}_t^{65} - \beta\mathrm{V}_t + \gamma 1_{\{b\mathrm{A}_t+c\mathrm{C}_t-d>0\}} + \delta \\ 2, & \mathrm{V}_t < -x\mathrm{C}_t - y\mathrm{I}_t^{65} + z \quad \mathrm{C}_t \geq \alpha\mathrm{I}_t^{65} - \beta\mathrm{V}_t + \gamma 1_{\{b\mathrm{A}_t+c\mathrm{C}_t-d>0\}} + \delta \\ 3, & \mathrm{V}_t \geq -x\mathrm{C}_t - y\mathrm{I}_t^{65} + z \quad \mathrm{C}_t < \alpha\mathrm{I}_t^{65} - \beta\mathrm{V}_t + \gamma 1_{\{b\mathrm{A}_t+c\mathrm{C}_t-d>0\}} + \delta \\ 4, & \mathrm{V}_t \geq -x\mathrm{C}_t - y\mathrm{I}_t^{65} + z \quad \mathrm{C}_t \geq \alpha\mathrm{I}_t^{65} - \beta\mathrm{V}_t + \gamma 1_{\{b\mathrm{A}_t+c\mathrm{C}_t-d>0\}} + \delta \end{cases}$$

$$\tau_{\mathrm{V}}^{(\zeta_t)} \triangleq \begin{cases} \tau_{\mathrm{V}}^{(1)}, & \zeta_t = 1 \text{ or } 2 \\ \tau_{\mathrm{V}}^{(2)}, & \zeta_t = 3 \text{ or } 4 \end{cases}$$

$$\tau_{\mathrm{C}}^{(\zeta_t)} \triangleq \begin{cases} \tau_{\mathrm{C}}^{(1)}, & \zeta_t = 1 \text{ or } 3 \\ \tau_{\mathrm{C}}^{(2)}, & \zeta_t = 2 \text{ or } 4 \end{cases} \tag{4.11}$$

Note that ζ_t is defined by the given parameters, I_t^{65} and S_t.

The particle smoother in section A.3.2 is completely specified by its transition (4.11) and emission (4.7) models. Suppose that we have drawn a bunch of sampled paths $\left\{\widetilde{\mathcal{S}}^{(l)}\right\}_{l=1}^{\mathrm{L}} = \left\{\widetilde{\mathcal{V}}^{(l)}, \widetilde{\mathcal{A}}^{(l)}, \widetilde{\mathcal{C}}^{(l)}\right\}_{l=1}^{\mathrm{L}}$. On the one hand, the covariance matrix Σ is easily updated by the following form:

$$\hat{\Sigma} = \frac{1}{\mathrm{L}\,(\mathrm{T}-1)} \sum_{l=1}^{\mathrm{L}} \sum_{t=1}^{\mathrm{T}-1} \sum_{k=1}^{4} \left(\widetilde{S}_{t+1}^{(n)} - \mathrm{W}_k \widetilde{S}_t^{(n)} - \mathrm{B}_k\right) \left(\widetilde{S}_{t+1}^{(n)} - \mathrm{W}_k \widetilde{S}_t^{(n)} - \mathrm{B}_k\right)^{\mathbb{T}} 1_{\left\{\widetilde{\zeta}_t^{(n)}=k\right\}} \tag{4.12}$$

For learning parameters, on the other hand, we have the following objective function:

$$f\left(\widetilde{\mathcal{S}}; \Theta\right) \triangleq -\frac{1}{2} \sum_{n=1}^{\mathrm{N}} \sum_{t=1}^{\mathrm{T}-1} \sum_{k=1}^{4} \left(\widetilde{\mathrm{S}}_{t+1}^{(n)} - \mathrm{W}_k \widetilde{\mathrm{S}}_t^{(n)} - \mathrm{B}_k\right)^{\mathbb{T}} \hat{\Sigma}^{-1} \left(\widetilde{\mathrm{S}}_{t+1}^{(n)} - \mathrm{W}_k \widetilde{\mathrm{S}}_t^{(n)} - \mathrm{B}_k\right) 1_{\left\{\widetilde{\zeta}_t^{(n)}=k\right\}} \tag{4.13}$$

We do not have any closed form for updating any of the parameters with (4.13): instead, we can adopt gradient descent methods, after regularizing the only nondifferentiable term: for a large $\lambda > 0$ (we use $\lambda = 1000$),

$$1_{\{b\mathrm{A}_t + c\mathrm{C}_t - d > 0\}} \approx \frac{1}{1 + \exp\left(-b\lambda \mathrm{A}_t - c\lambda \mathrm{C}_t + d\lambda\right)} \tag{4.14}$$

Note that each partial derivative of (4.13) is given as follows:

$$\frac{\partial f}{\partial \theta}\left(\widetilde{\mathcal{S}}; \Theta\right) = \sum_{n=1}^{\mathrm{N}} \sum_{t=1}^{\mathrm{T}-1} \sum_{k=1}^{4} \left(\widetilde{\mathrm{S}}_{t+1}^{(n)} - \mathrm{W}_k \widetilde{\mathrm{S}}_t^{(n)} - \mathrm{B}_k\right)^{\mathbb{T}} \hat{\Sigma}^{-1} \left(\frac{\partial \mathrm{W}_k}{\partial \theta} \widetilde{\mathrm{S}}_t^{(n)} + \frac{\partial \mathrm{B}_k}{\partial \theta}\right) 1_{\left\{\widetilde{\zeta}_t^{(n)}=k\right\}} \tag{4.15}$$

By plugging the approximation (4.14) in the objective function (4.13), parameters are updated. Now we are prepared. Algorithm 4 summarizes the inference procedure.

4.3.2 Gaussian Process State-space Model

Although we already covered the GPST briefly in section B.4, here we give a more detailed inference procedure. Let us first define the notations as follows:

Algorithm 4: Particle Smoother Atmospheric CO_2 Reconstruction

```
initialize parameters and covariance matrix
while converging do
    sample dimensionless index paths by the particle smoother algorithm
    update the covariance matrix given parameters and sampled paths
    update the parameters given covariance matrix and sampled paths
end
restore atmospheric CO2 from the associated dimensionless indices
return a set of sampled paths of CO2
```

- $\mathcal{T} = \mathrm{T}_{1:\mathrm{N}}$: ages of observations.
- $\mathcal{X} = \mathrm{X}_{1:\mathrm{N}}$: atmospheric CO_2 at $\mathcal{T}$ to infer.
- $\mathcal{Y} = \mathrm{Y}_{1:\mathrm{N}}$: $\delta^{11}\mathrm{B}$ at $\mathcal{T}$.
- $\mathcal{V} = \mathrm{V}_{1:\mathrm{N}}$: ice volumes at $\mathcal{T}$ somehow correlated with $\mathcal{X}$.

Then, we have the following models on them. First, what we model as a GP is the latent state $\mathcal{X}$ over $\mathcal{T}$: recall that Matérn kernels are corresponding to the continuous AR models (sections B.1.1 and B.4).

$$p\left(\mathcal{X}|\mathcal{T}\right) = \mathcal{N}\left(\mathcal{X}\middle|\underset{\sim}{\mu}_{\mathcal{T}}, \mathbb{K}_{\mathcal{T}\mathcal{T}}\right) \tag{4.16}$$

Secondly, to deal with outliers, we consider a Student's t-distribution with 6 as the degree of freedom for the emission model of $\delta^{11}\mathrm{B}$ given $\mathcal{X}$ and $\mathcal{T}$: the mean α and variance τ are the same as those of (4.6).

$$\begin{aligned} p\left(\mathcal{Y}|\mathcal{X},\mathcal{T}\right) &= \prod_{n=1}^{\mathrm{N}} \mathcal{T}\left(\mathrm{Y}_n|\alpha\left(\mathrm{X}_n\right), \tau\left(\mathrm{X}_n\right), 6\right) \\ \alpha\left(\mathrm{X}_n\right) &= 30.3 + 0.0552\mathrm{X}_n - 3.80\log\left(\mathrm{X}_n + 8.76\right) \\ \tau\left(\mathrm{X}_n\right) &= \exp\left(-3.32 + 0.0392\mathrm{X}_n\right) \end{aligned} \tag{4.17}$$

Lastly, a Gaussian mixture model [Murphy (2012)] is considered for the emission model of ice

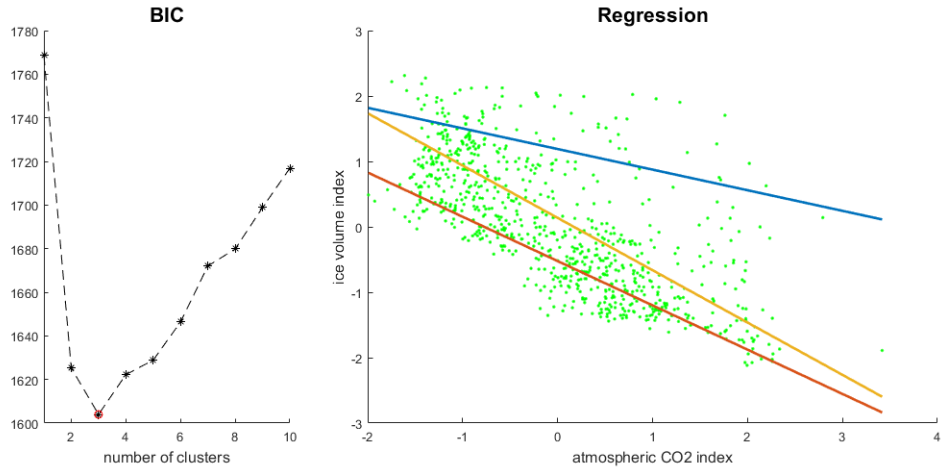


Figure 4.4: (Left) The BIC curve that selects 3 as the number of clusters and (Right) atmospheric CO_2 and ice volume index pairs (green) and the regressions of the clusters (in each color).

volumes given $\mathcal{X}$: figure 4.4 visualizes the training data and learned model.

$$p\left(\mathcal{V}|\mathcal{X}\right) = \prod_{n=1}^{\mathrm{N}} \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n \middle| a_k \mathrm{X}_n + b_k, \omega_k^2\right) \tag{4.18}$$

In the GP regression model, the goal is to obtain the marginalized distribution of the proxy given the associated input. Here, the model aims to infer $\mathcal{X}$ given $\mathcal{T}$, $\mathcal{V}$ and $\mathcal{Y}$. Note that the likelihoods (4.17) and (4.18) are not Gaussian, thus the posterior of $\mathcal{X}$ cannot be given as a known distribution.

A plausible approach is sampling $\mathcal{X}$ from the posterior $p\left(\mathcal{X}|\mathcal{T},\mathcal{V},\mathcal{Y}\right)$ which is proporational to $p\left(\mathcal{Y}|\mathcal{X},\mathcal{T}\right)p\left(\mathcal{V}|\mathcal{X}\right)p\left(\mathcal{X}|\mathcal{T}\right)$ by a Markov chain Monte Carlo (MCMC) (section A.4). It is, however, limited because 1) $\mathcal{X}$ tends to be highly correlated one another so it would require a lot of iteration until reaching a burn-in period and 2) computing $p\left(\mathcal{X}|\mathcal{T}\right)$ (or $p\left(\mathrm{X}_n|\mathcal{X}\setminus \mathrm{X}_n,\mathcal{T}\right)$) becomes intractable quickly as the size of data N grows.

Instead, we consider the following variational method. Consider $q\left(\mathcal{X}|\Theta\right)$ as an approximation of $p\left(\mathcal{X}|\mathcal{T},\mathcal{V},\mathcal{Y}\right)$ defined as follows: for $\mu = \left(\mu_1,\cdots,\mu_{\mathrm{N}}\right)^{\mathbb{T}}$ and $\Psi = diag\left(\psi_1^2,\cdots,\psi_{\mathrm{N}}^2\right)$,

$$q\left(\mathcal{X}|\Theta\right) = \mathcal{N}\left(\mathcal{X}|\mu,\Psi\right) = \prod_{n=1}^{\mathrm{N}} \mathcal{N}\left(X_n \middle| \mu_n, \psi_n^2\right) \tag{4.19}$$

A subsequent issue is about how to estimate $\Theta = \theta_{1:\mathrm{N}} = \left\{\mu_{1:\mathrm{N}}, \psi_{1:\mathrm{N}}\right\}$. Let us consider the following Kullback-Leibler (KL) divergence:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathcal{T},\mathcal{V},\mathcal{Y}\right)\right) \\
&= \log p\left(\mathcal{V},\mathcal{Y}|\mathcal{T}\right) + \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathcal{T}\right)\right) - \mathbb{E}_q\left[\log p\left(\mathcal{Y}|\mathcal{X},\mathcal{T}\right) + \log p\left(\mathcal{V}|\mathcal{X}\right)\right] \\
&= \log p\left(\mathcal{V},\mathcal{Y}|\mathcal{T}\right) + \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathcal{T}\right)\right) - \sum_{n=1}^{\mathrm{N}} \mathbb{E}_q\left[\log \mathcal{T}\left(\mathrm{Y}_n|\alpha\left(\mathrm{X}_n\right),\tau\left(\mathrm{X}_n\right),6\right)\right] \\
&- \sum_{n=1}^{\mathrm{N}} \mathbb{E}_q\left[\log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n\middle|a_k\mathrm{X}_n + b_k, \omega_k^2\right)\right]
\end{aligned} \tag{4.20}$$

Since KL divergences are nonnegative, we have the following inequality giving the evidence lower bound (ELBO) to the log-marginal likelihood (log-ML) $\log p\left(\mathcal{V},\mathcal{Y}|\mathcal{T}\right)$:

$$\begin{aligned}
\log p\left(\mathcal{V},\mathcal{Y}|\mathcal{T}\right) \geq &- \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathcal{T}\right)\right) + \sum_{n=1}^{\mathrm{N}} \mathbb{E}_q\left[\log \mathcal{T}\left(\mathrm{Y}_n|\alpha\left(\mathrm{X}_n\right),\tau\left(\mathrm{X}_n\right),6\right)\right] \\
&+ \sum_{n=1}^{\mathrm{N}} \mathbb{E}_q\left[\log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n\middle|a_k\mathrm{X}_n + b_k, \omega_k^2\right)\right]
\end{aligned} \tag{4.21}$$

Because both $p\left(\mathcal{X}|\mathcal{T}\right)$ and $q\left(\mathcal{X}|\Theta\right)$ are Gaussian, we can represent $\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathcal{T}\right)\right)$ in a closed form as follows:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\Theta\right)||p\left(\cdot|\mathcal{T}\right)\right) \\
&= \frac{1}{2}\left(\mathrm{trace}\left(\mathbb{K}_{\mathcal{T}\mathcal{T}}^{-1}\Psi\right) + \left(\mu - \underline{\mu}_{\mathcal{T}}\right)^{\mathbb{T}} \mathbb{K}_{\mathcal{T}\mathcal{T}}^{-1}\left(\mu - \underline{\mu}_{\mathcal{T}}\right) - 1 + \log\frac{|\mathbb{K}_{\mathcal{T}\mathcal{T}}|}{|\Psi|}\right)
\end{aligned} \tag{4.22}$$

The last term cannot be explicitly computed. One possible treatment is to use the following lower bound holds for each n by Jensen's inequality:

$$\begin{aligned}
&\mathbb{E}_q\left[\log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n\middle|a_k\mathrm{X}_n + b_k, \omega_k^2\right)\right] \geq \mathbb{E}_q\left[\sum_{k=1}^{\mathrm{K}} p_k \log \mathcal{N}\left(\mathrm{V}_n\middle|a_k\mathrm{X}_n + b_k, \omega_k^2\right)\right] \\
&= \sum_{k=1}^{\mathrm{K}} p_k\left(\log \mathcal{N}\left(\mathrm{V}_n\middle|a_k\mu_n + b_k, \omega_k^2\right) - \frac{1}{2}a_k^2\omega_k^{-2}\psi_n^2\right)
\end{aligned} \tag{4.23}$$

However, it turns out that the gap in (4.23) is too big to guarantee a good variational approximation after running the real data. Therefore, a Monte Carlo integration should be considered instead. To reduce the variance, let us consider the reparameterization trick, as discussed in sections B.2.3 and B.4: for each n, let $\varepsilon_n \sim \mathcal{N}(0, 1^2)$. Then, the expectation can be represented in the following form:

$$\begin{aligned} &\mathbb{E}_q \left[\log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n \middle| a_k \mathrm{X}_n + b_k, \omega_k^2\right) \right] \\ &= \mathbb{E}_{\varepsilon_n} \left[\log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n \middle| a_k \left(\mu_n + \psi_n \varepsilon_n\right) + b_k, \omega_k^2\right) \right] \end{aligned} \tag{4.24}$$

Therefore, its gradient with respect to θ_n is given as follows: in practice, we approximate it with the Monte Carlo integration.

$$\begin{aligned} &\nabla_{\theta_n} \mathbb{E}_q \left[\log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n \middle| a_k \mathrm{X}_n + b_k, \omega_k^2\right) \right] \\ &= \mathbb{E}_{\varepsilon_n} \left[\nabla_{\theta_n} \log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(\mathrm{V}_n \middle| a_k \left(\mu_n + \psi_n \varepsilon_n\right) + b_k, \omega_k^2\right) \right] \end{aligned} \tag{4.25}$$

It is also not possible to express each $\mathbb{E}_q \left[\log \mathcal{T}\left(\mathrm{Y}_n | \alpha\left(\mathrm{X}_n\right), \tau\left(\mathrm{X}_n\right), 6\right)\right]$ explicitly. As similar with above, we have the following:

$$\begin{aligned} \mathbb{E}_q \left[\log \mathcal{T}\left(\mathrm{Y}_n | \alpha\left(\mathrm{X}_n\right), \tau\left(\mathrm{X}_n\right), 6\right)\right] &= \mathbb{E}_{\varepsilon_n} \left[\log \mathcal{T}\left(\mathrm{Y}_n \middle| \alpha\left(\mu_n + \psi_n \varepsilon_n\right), \tau\left(\mu_n + \psi_n \varepsilon_n\right), 6\right)\right] \\ \nabla_{\theta_n} \mathbb{E}_q \left[\log \mathcal{T}\left(\mathrm{Y}_n | \alpha\left(\mathrm{X}_n\right), \tau\left(\mathrm{X}_n\right), 6\right)\right] &= \mathbb{E}_{\varepsilon_n} \left[\nabla_{\theta_n} \log \mathcal{T}\left(\mathrm{Y}_n \middle| \alpha\left(\mu_n + \psi_n \varepsilon_n\right), \tau\left(\mu_n + \psi_n \varepsilon_n\right), 6\right)\right] \end{aligned} \tag{4.26}$$

Therefore, we can run a gradient ascend on the lower bound in (4.21) to variationally maximize $p(\mathcal{V}, \mathcal{Y} | \mathcal{T})$ under mild regularity conditions, including that α and τ are partially differentiable with respect to the first coordinates. Note that it also becomes intractable as N grows because (4.22) is involved in computing $\mathbb{K}_{\mathcal{T}\mathcal{T}}^{-1}$. This can be dealt with by using small batches instead of the full data in feeding in the gradient ascent or considering another variational approximation on $p(\mathcal{X} | \mathcal{T}, \mathcal{V}, \mathcal{Y})$.

Suppose that Θ and $\mathbb{K}$ are estimated by so. Then, for an arbitrary age t, the associated atmospheric CO_2 x has the following posterior predictive:

$$\begin{aligned} p\left(x|t,\mathcal{T},\mathcal{Y},\mathcal{V}\right) &= \int p\left(x|t,\mathcal{X},\mathcal{T}\right) p\left(\mathcal{X}|\mathcal{T},\mathcal{Y},\mathcal{V}\right) d\mathcal{X} \\ &\approx \int p\left(x|t,\mathcal{X},\mathcal{T}\right) q\left(\mathcal{X}|\Theta\right) d\mathcal{X} = \mathcal{N}\left(x|\rho\left(t\right),\delta\left(t\right)\right) \\ \rho\left(t\right) &= \mathbb{K}_{t\mathcal{T}}\mathbb{K}_{\mathcal{T}\mathcal{T}}^{-1}\left(\underset{\sim}{\mu} - \underset{\sim\mathcal{T}}{\mu}\right) + \underset{\sim t}{\mu} \\ \delta\left(t\right) &= \mathbb{K}_{tt} - \mathbb{K}_{t\mathcal{T}}\mathbb{K}_{\mathcal{T}\mathcal{T}}^{-1}\mathbb{K}_{\mathcal{T}t} + \mathbb{K}_{t\mathcal{T}}\mathbb{K}_{\mathcal{T}\mathcal{T}}^{-1}\Psi\mathbb{K}_{\mathcal{T}\mathcal{T}}^{-1}\mathbb{K}_{\mathcal{T}t} \end{aligned} \tag{4.27}$$

However, we can do slightly better than this if ice volumes are given at any ages. Let v be that value at t. Then, the density of v given x is $\sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}\left(v|a_k x + b_k, \omega_k^2\right)$, which can be plugged in as a likelihood, so we have the following posterior of x, with (4.27) as the prior:

$$\begin{aligned} p\left(x|t,v,\mathcal{T},\mathcal{Y},\mathcal{V}\right) &= \sum_{k=1}^{\mathrm{K}} \bar{p}_k\left(t,v\right)\mathcal{N}\left(x\middle|\overline{\mu}_k\left(t,v\right),\overline{\nu}_k\left(t,v\right)\right) \\ &\propto p\left(x|t,\mathcal{T},\mathcal{Y},\mathcal{V}\right)\sum_{k=1}^{\mathrm{K}} p_k\mathcal{N}\left(v\middle|a_k x + b_k,\omega_k^2\right) \\ \bar{p}_k\left(t,v\right) &= \frac{p_k\mathcal{N}\left(v|a_k\rho\left(t\right)+b_k,\omega_k^2+a_k^2\delta\left(t\right)\right)}{\sum_{l=1}^{\mathrm{K}} p_l\mathcal{N}\left(v|a_l\rho\left(t\right)+b_l,\omega_l^2+a_l^2\delta\left(t\right)\right)} \\ \overline{\mu}_k\left(t,v\right) &= \left(\delta\left(t\right)^{-1}+a_k^2\omega_k^{-2}\right)^{-1}\left(a_k\omega_k^{-2}\left(v-b_k\right)+\delta\left(t\right)^{-1}\rho\left(t\right)\right) \\ \overline{\nu}_k\left(t,v\right) &= \left(\delta\left(t\right)^{-1}+a_k^2\omega_k^{-2}\right)^{-1} \end{aligned} \tag{4.28}$$

To get the confidence bands with (4.28), note that, for any η, we have the following:

$$\begin{aligned} \mathbb{P}\left(x \geq \eta|t,v,\mathcal{T},\mathcal{Y},\mathcal{V}\right) &= \sum_{k=1}^{\mathrm{K}} \bar{p}_k\left(t,v\right)\left(1-\Phi\left(\frac{\eta-\overline{\mu}_k\left(t,v\right)}{\sqrt{\overline{\nu}_k\left(t,v\right)}}\right)\right) \\ \mathbb{P}\left(x < \eta|t,v,\mathcal{T},\mathcal{Y},\mathcal{V}\right) &= \sum_{k=1}^{\mathrm{K}} \bar{p}_k\left(t,v\right)\Phi\left(\frac{\eta-\overline{\mu}_k\left(t,v\right)}{\sqrt{\overline{\nu}_k\left(t,v\right)}}\right) \end{aligned} \tag{4.29}$$

Now we just need to get numerical η's to make the upper and lower term in (4.29) $\alpha/2$ for obtaining the $100\left(1-\alpha\right)\%$ confidence level.

Now we are again prepared. Algorithm 5 summarizes the inference procedure.

Algorithm 5: Gaussian Process State-space Model for Atmospheric CO_2 Reconstruction

```
initialize kernel hyperparameters and variational parameters
while converging do
    tune kernel hyperparameters
    learn variational parameters
end
compute and return the posterior predictive mean and variance functions
```

4.4 Results and Discussion

We tried two models for reconstructing atmospheric CO_2, one is the particle smoother and the other is the GPST. Each set of results are visualized in the corresponding figure that has three features. The black dashed curve and shaded region are that of reconstructed atmospheric CO_2 (the median of the sampled paths) and its 95% confidence band, respectively. The red curve indicates the "true" atmospheric CO_2 (experimental time series) from the Antarctic ice core records. The blue bars are the lower-upper intervals of the published CO_2 of [Dyez et al. (2018)].

4.4.1 Particle Smoother

We set the number of particles to be 200 and ran 100 iterations, where only the covariance matrix was supposed to be learned for the first 30 iterations. The learned covariance matrix and parameters are given in (4.30) and figure 4.5.

$$\hat{\Sigma} = \begin{pmatrix} 0.0464 & -0.0322 & -0.0298 \\ -0.0322 & 1.7145 & 0.0511 \\ -0.0298 & 0.0511 & 0.1850 \end{pmatrix} \tag{4.30}$$

Figure 4.6 shows three panels, one for each kind of hidden states. It is clear that the 95% confidence band of the reconstructed extent of the Antarctic ice sheet indices is far broader than

	Initial Value	Learned Parameter
α	0.237	-0.0618
β	0.793	1.0176
γ	1.955	0.0072
δ	0.146	1.1894
x	0.669	0.4840
y	0.527	0.8121
z	0.761	0.7416
b	0.936	0.8466
c	0.533	0.6288
d	0.069	-0.0779
$\tau_{\mathrm{C}}^{(1)}$	2.793	11.5089
$\tau_{\mathrm{C}}^{(2)}$	8.414	1.6338
τ_{A}	10.266	1.9672
$\tau_{\mathrm{V}}^{(1)}$	11.425	10.1024
$\tau_{\mathrm{V}}^{(2)}$	2.325	1.6374

Figure 4.5: Initial and learned values of the parameters in the particle smoother model. The initial values were selected from [Garcia-Olivares & Herrero (2013)].

the other indices because neither benthic δ^{18}O or δ^{11}B proxy is paired with it. The same logic is applied to the narrow band of the reconstructed ice volume indices because benthic δ^{18}O observations are plentiful compared to δ^{11}B. Compared to the lower-upper intervals of the published ones, our 95% confidence band of the reconstructed atmospheric CO_2 is narrower thus more certain: it makes sense because we used not only δ^{11}B but also benthic δ^{18}O for the inference. Both reconstructions mostly cover the experimental time series as the "true" atmospheric CO_2 curve from the Antarctic ice core records, but the deviations appeared at 190-230 kiloyears: this could imply the model misspecification because the ice volume indices are also deviated from its experimental time series at the same range of ages.

Note that all the results become plausible under the assumption that the transition model is correctly specified and stationary over time. In fact, such assumptions are too strong, as criticized by a few attendees in the 2nd Workshop on Cenozoic pCO_2 Reconstruction. It means that the learned parameters cannot be reused for reconstructing the atmospheric CO_2 in different age ranges. Also, to make the data fit into the model, we adjusted their ages to the grid of every 1 kiloyear, so that the information loss is structurally inevitable.

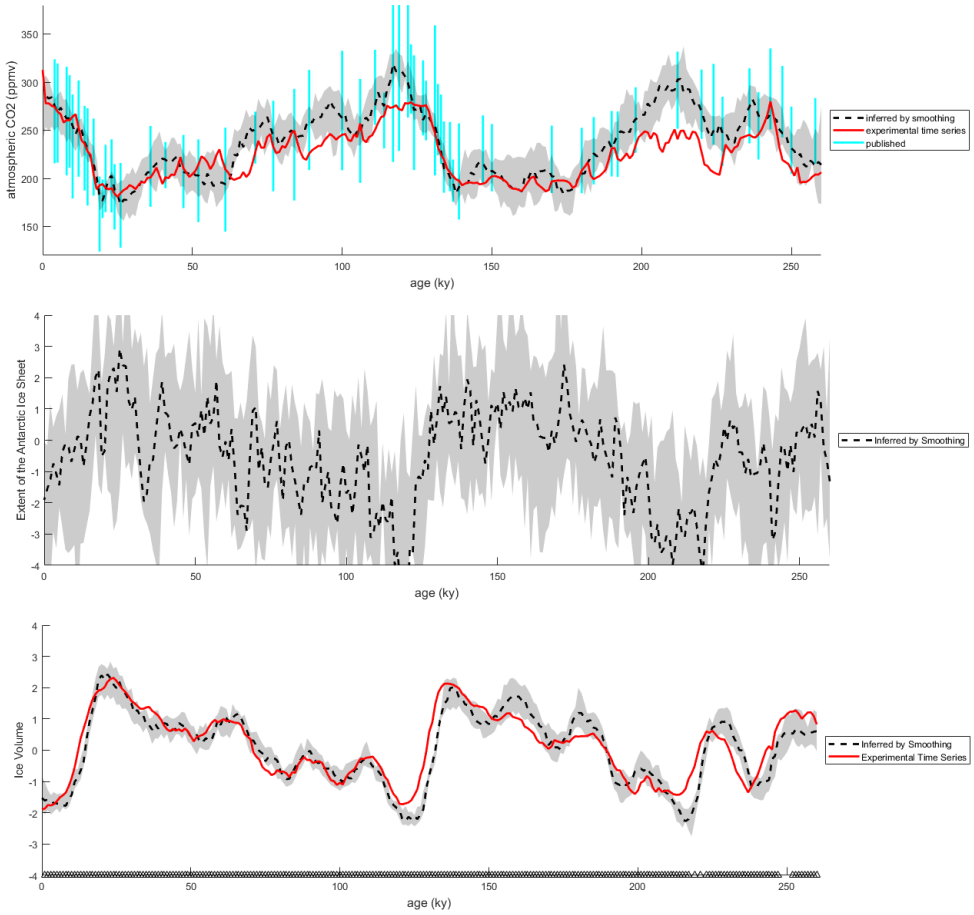


Figure 4.6: Results from the particle smoother. (Top) Reconstructed atmospheric CO_2 with the experimental time series and published lower-upper bands. (Middle) Reconstructed extent of the Antarctic ice sheet indices. (Bottom) Reconstructed ice volume indices with the experimental time series from [Spratt & Lisiecki (2016)].

4.4.2 Gaussian Process State-space Model

Figures 4.7 to 4.10 visualize the results for the SE kernel, Matérn kernel with $d = 5/2$ (M25), Matérn kernel with $d = 3/2$ (M15) and OU kernel, respectively. Each figure consists of four panels: the first and second panel show the inference by the full model and the third and fourth panels present the results not using ice volumes as the proxy. Definitely the former gives narrower confidence bands to the latter because of more proxies. The second and fourth panels span up to 800 kiloyears while the first and third ones are up to 260 kiloyears to be coupled with the particle smoother results in section 4.4.1. The reconstructed atmospheric CO_2 series is now smoother and mostly covers the experimental time series, and the sensitiveness to the data is in order of SE, M25, M15 and OU, as expected by the kernel characteristics.

The nonparametric aspect of GPST means that the transition model is agnostic and dependent on the data directly. Also, it does not need regularly arranged data over ages. Thus, the model is now free from the strong assumptions on the transition model in the particle smoother (section 4.4.1). However, the inference is now not exact but variational. It is possible to overcome such aspect in particle smoother by increasing the number of particles, but GPST needs to adopt more general (complicated) parametric variational distributions for better approximation, that potentially leads to overfitting.

4.5 Future Work

As mentioned earlier, all the results are quite preliminary: each method adopts emission models that are constructed under very strong assumptions. Published ages and training data are given as true values and observational models do not reflect any particular data-specific characteristics but are just reduced to some simple parametric models. The dataset is also too scarce: there are less than 100 boron proxies up to 800 kiloyears. To overcome the hypothetical aspects stemming from such limits, the interdisciplinary research should be considered. With a help of the experts of the related fields, more reasonable transitional and observational models and more objective

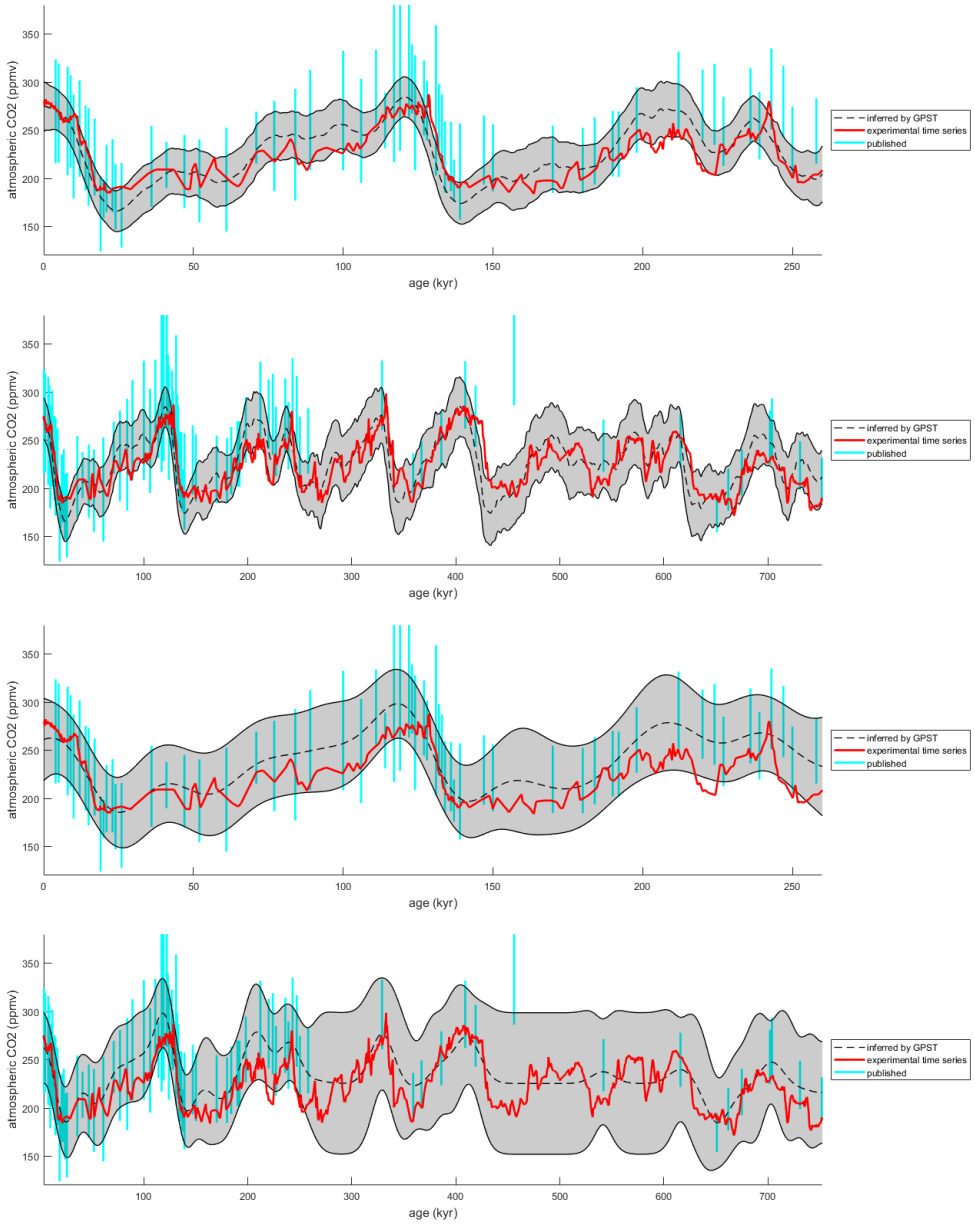


Figure 4.7: Results from the Gaussian process state-space model with the SE kernel.

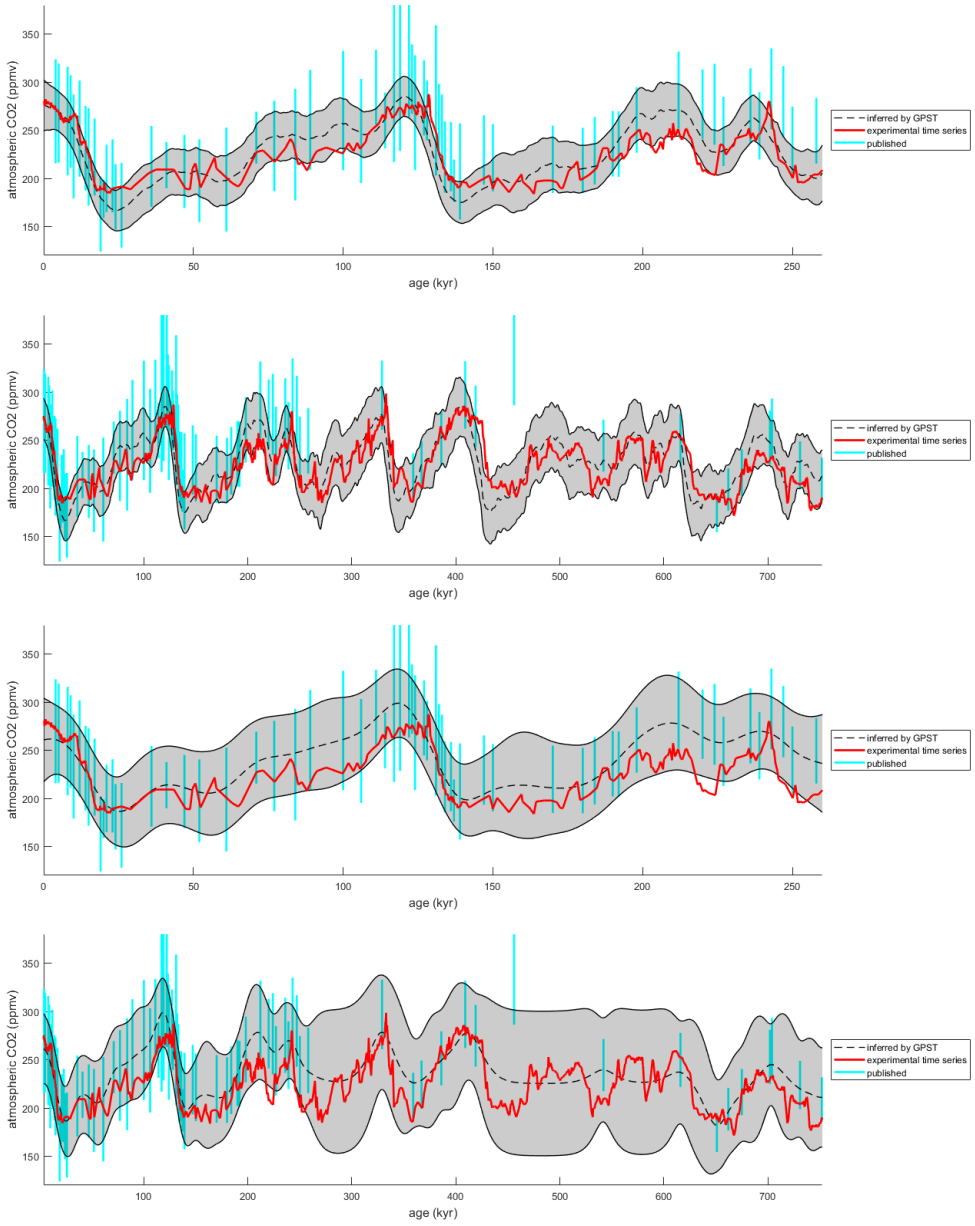


Figure 4.8: Results from the Gaussian process state-space model with the M25 kernel.

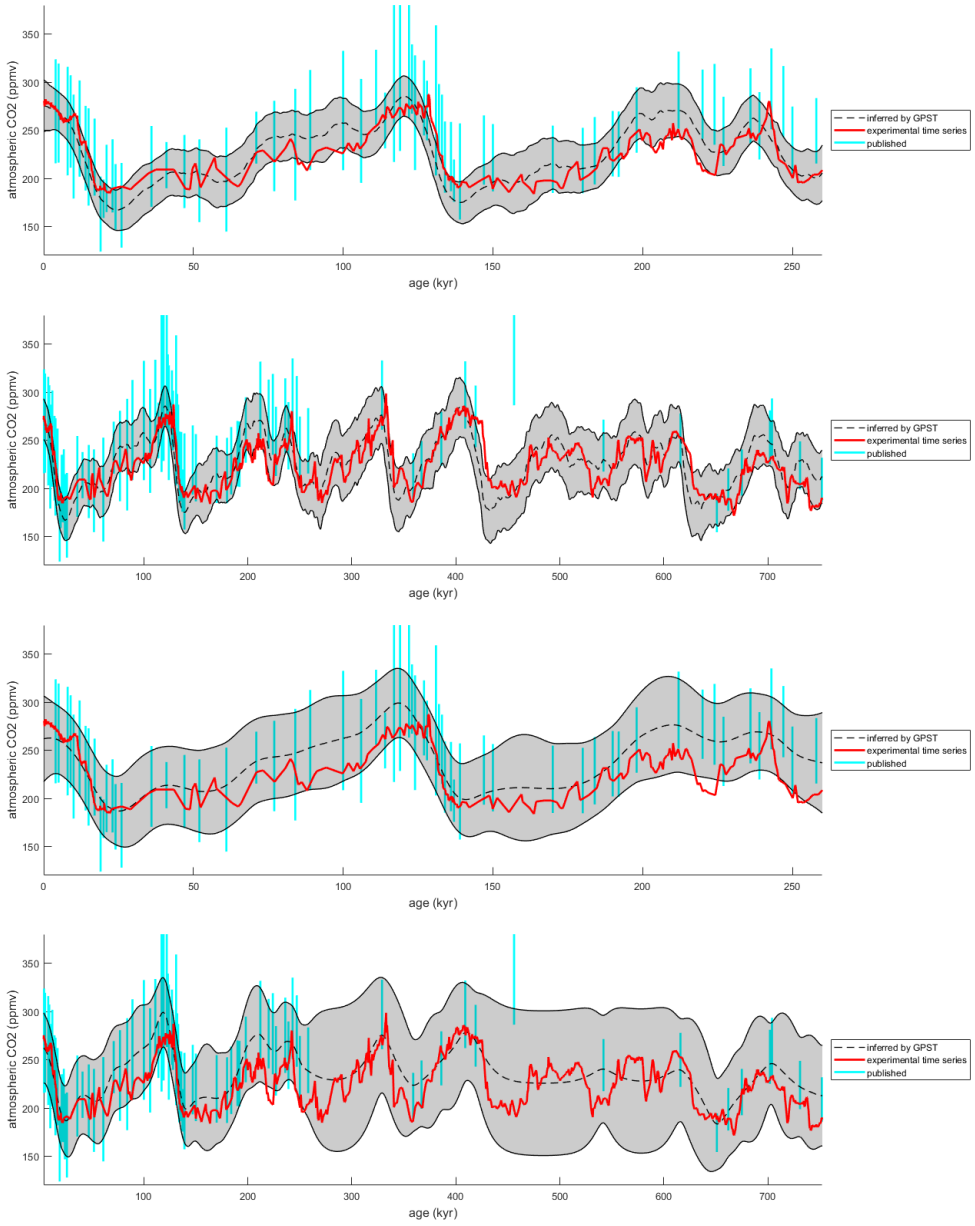


Figure 4.9: Results from the Gaussian process state-space model with the M15 kernel.

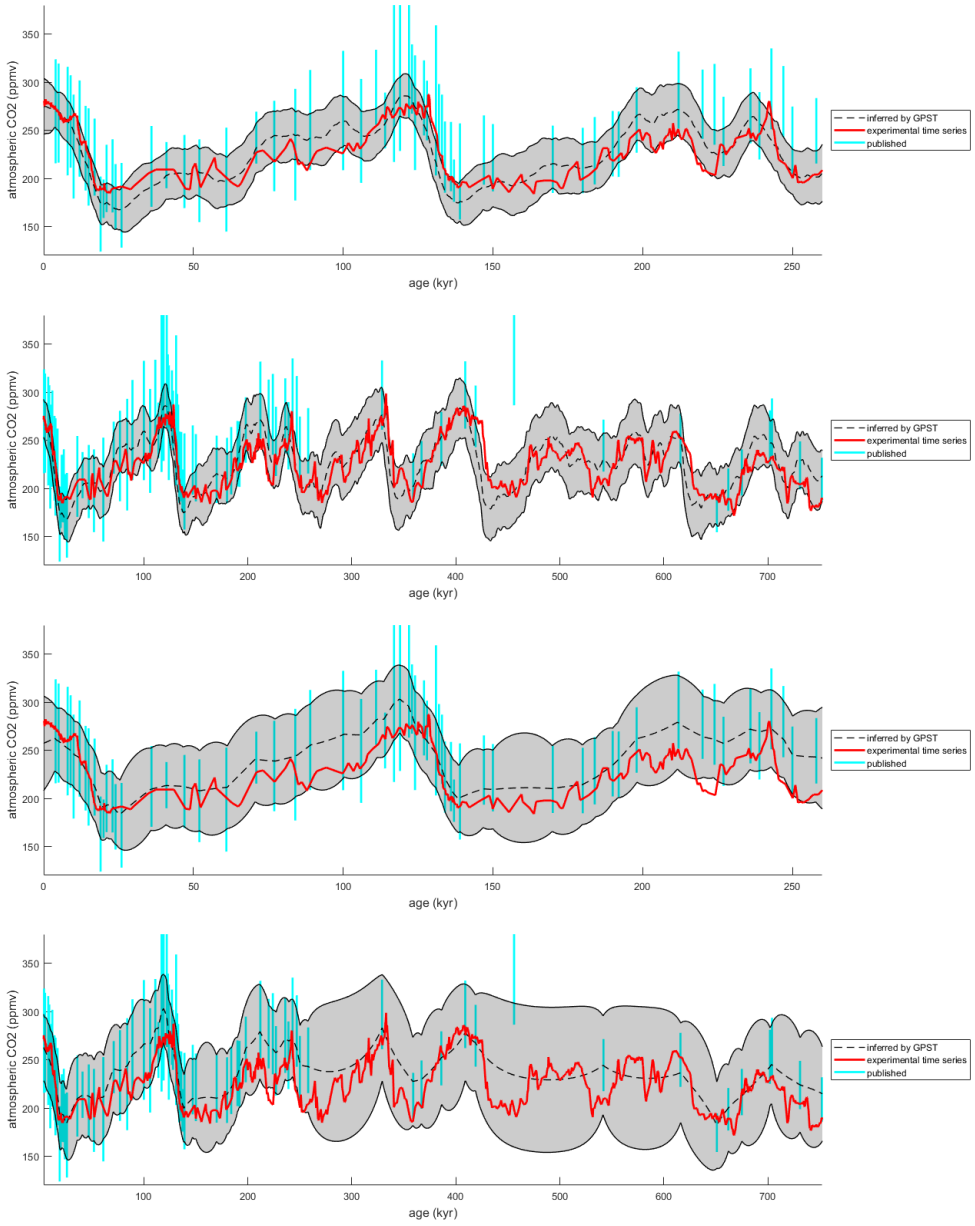


Figure 4.10: Results from the Gaussian process state-space model with the OU kernel.

training data will be available.

We tried four Matérn kernels in GPST modelling. As discussed in chapter B, the suboptimality can be overcome by adding multiple hidden layers (deep Gaussian process (GP) in section B.6) or adopting more complicated kernels. Though applying deep GP models to this problem is also intriguing, using weighted mixtures of kernels with periodic kernels (section B.1.1) would be an easy solution to deal with this problem, due to the apparently periodic pattern of the atmospheric CO_2 over time. More complicated transition models can be chosen in the particle smoother but the time complexity and curse of dimensionality on the particle smoother may preclude such attempts.

As discussed in chapter 2 about how the direct and indirect proxies are complementary, using various proxies that are correlated with the target ocean and climate events would be a good solution for overcoming the limited data availability. Figures 4.7 to 4.10 show the advantage of using multiple proxies vividly. Here we used benthic $\delta^{18}O$ and sea level data (that are transformed into the ice volumes) as such proxies, but did not use proxies for the other events, such as salinity and sea surface temperatures (SSTs), that also affect the CO_2 level. Constructing comprehensive models could be regarded as another future work, just as [Bowen et al. (2020)] did for reconstructing SSTs.

5

Conclusion

Here we presented some applications of the Bayesian statistical learning in connection with the fields of geoscience, paleoceanography and paleoclimatology. Various state-space models and Gaussian process models were chosen to utilize the proxies for reconstructing the relevant past ocean and climate events and for assigning ages to them. New alignment algorithm, improved calibration models, novel methodologies and promising preliminary works have been covered together with the results from the real data throughout the dissertation. The theoretical backgrounds of state-space models and Gaussian process models are extensively covered in Appendix.

My graduate research topics about paleoceanography and paleoclimatology were mainly covered through chapters 2 to 4. Chapter 2 is about the benthic δ^{18}O stack construction method,

a special case of SA-GPR algorithm for the multiple signal alignments, introduced in a submitted paper of mine, for inferring ages of cores. It is a motif-finding algorithm thus consisting of the core (signal) alignment algorithm and the stack (motif) construction algorithm. Each core is aligned to the target stack by a state-space model that is a hybrid of the particle smoother for initializing sampled age paths and Metropolis-Hastings algorithm for the refinements. The stack construction algorithm starts with utilizing the homoscedastic Gaussian process regression method that constructs a sample-specific model for each sampled age path. Then a core-specific stack is constructed from the affiliated sample-specific models by the model reduction method based on the moment-matching. Finally, a local stack is obtained by the variational method. This algorithm is for the cores in the same homogeneous set, which share the same local effect. In future, we expect to refine the model with differentiable transition models that allow the Hamiltonian Monte Carlo algorithm for efficient sampling, to extend it to the hierarchical stack construction method for handling multiple homogeneous sets simultaneously, and to add more age proxies for enhancing the age inference.

Chapter 3 covered the calibration model construction algorithms based on the Gaussian process regression, that extends the heteroscedastic Gaussian process regression model in a conference paper of mine. Both homoscedastic and heteroscedastic Gaussian process regression models with parametric mean prior functions were tried on the alkenone and TEX86 proxy data used for reconstructing sea surface temperatures individually. To deal with the large datasets, the variational extension based on the variational free energy was added to the exact models. For verifying the multimodality of the data, a novel multimodal Gaussian process regression algorithm that can be learned by the explicit expectation-maximization algorithm was designed and also applied for the datasets. In future, we will process the outliers more carefully by classifying them into two categories, one is the set of noninformative noises and the other is for the systematically contaminated data. For the latter case, we are working with some experts in the alkenone proxy for dealing with lab-specific biases. Constructing site-specific calibration models is another intriguing goal: the regression models are then depending not only on the sea surface temperatures but also on

the query location. Applying deep Gaussian process models and designing Bayesian inference models for reconstructing sea surface temperatures from the calibration models will follow.

Chapter 4 was consisting of two preliminary works for reconstructing atmospheric CO_2 by using the boron proxy. Two state-space models were considered: one was the particle smoother with a parametric transition model on multiple correlated climate events and the other was a nonparametric modelling, the Gaussian process state-space model. By plugging the pre-trained emission models in those models and using more proxies such as the benthic $\delta^{18}O$ and historical sea level, we obtained promising inference results that are mostly consistent with the "true" atmospheric CO_2 from the Antarctic ice core records. Four Matérn kernels that correspond to the continuous autoregressive models were tried. Due to the data scarcity, using boron proxy only did not result in a good inference with the Gaussian process state-space model. All inferences in this chapter were, however, limited due to the plain parametric observational models. More reliable and reasonable results will be available by collaborating with the experts in the pCO_2 community. Designing more detailed kernels for the Gaussian process state-space models is another way of increasing the opportunity. Clearly, extending the model to taking more proxies is helpful for the inference, no doubt.

Appendix A and B covered the base knowledge of probabilistic state-space models and Gaussian processes, respectively. The reasons why I organized such knowledge in separate chapters are not only that some tools were repeatedly appeared throughout my research topics but also that I wanted my dissertation to be regarded as a manual for the beginners who hope to know more about the statistical models that can be applied to their data which could be essentially similar to ours. Appendix A covered HMMs as the discrete state-space models, Kalman filter and smoother as the continuous models that can be inferred exactly, particle filter and smoother as the variational methods for more general settings, mixture Kalman filter as an example of the Rao-Blackwellized particle filter for reducing variances, and some of the popular Markov-chain Monte Carlo algorithms as the alternatives for the particle filter and smoother of which performance is dependent on the number of particles. Appendix B was designed for introducing Gaus-

sian process models. The nonparametric characteristic of Gaussian processes leads the affiliated continuous models more flexible than the parametric models, thus allows us to apply them for various purposes, such as regression, classification, state-space models and dimensionality reduction. Gaussian process regression models return the explicit posterior predictive distribution. Gaussian process classification methods reflect the density of data. Gaussian process state-space models can handle the inputs that are gathered irregularly without adjustments. The Gaussian process latent variable model is a nonparametric dimensionality reduction. Gaussian processes have two drawbacks: one is that the obtained models are usually suboptimal due to the misspecified kernels. The other is about the time complexity that precludes the exact inference quickly as the size of training data increases. These two weaknesses can be addressed by the deep Gaussian process models and variational inference such as variational free energy.

For the last decades, machine learning community has mainly focused on processing and handling large datasets with complicated parametric models such as deep neural networks relying on the exponential growth in computing power. These innovations are clearly helpful for the researches in certain areas where large data can be easily gathered or generated at cheap prices in the reality or laboratories. In contrast, some other fields, that are hard, or even impossible, to gather or generate data as much as being wanted, have been relatively alienated. Parametric models that are correctly specified by theory are often adopted for dealing with such kinds of datasets, but what if the dataset is out of any good model as well? State-space models and the Gaussian process are for dealing with such datasets. Both methods give certain structures to the hidden variables. While a state-space model depends on the parametric transition model, the Gaussian process assumes a nonparametric prior on the hidden variables. The inference is then done either exactly or variationally. These Bayesian methods lead to utilize the limited data more efficiently. One more advantage is that these tools also access to the large datasets as well. State-space models have been frequently adopted in geosciences, especially (nonlinear) geophysics, but the use of Gaussian process models is still limited. I hope that my dissertation will lead the geoscience community to pay more attention to the Gaussian process modelling.

A

State-space Models

Let us begin with a simple example of the Bayesian inference that consists of a set of data $\mathcal{Y} = \{\mathrm{Y}_n\}_{n=1}^{\mathrm{N}}$, parametric model p, a finite-dimensional parameter θ and their prior π. Under the assumption that the data Y_n's are independent and identically distributed in (A.1), we have the following formulation:

$$p\left(\mathcal{Y}|\theta\right) = \prod_{n=1}^{\mathrm{N}} p\left(\mathrm{Y}_n|\theta\right) \tag{A.1}$$

$$p\left(\theta|\mathcal{Y}\right) = \frac{\pi\left(\theta\right) p\left(\mathcal{Y}|\theta\right)}{p\left(\mathcal{Y}\right)} \propto \pi\left(\theta\right) \prod_{n=1}^{\mathrm{N}} p\left(\mathrm{Y}_n|\theta\right) \tag{A.2}$$

In other words, the posterior of θ given $\mathcal{Y}$ in (A.2) is proportional to the product of $p\left(\mathrm{Y}_n|\theta\right)$'s

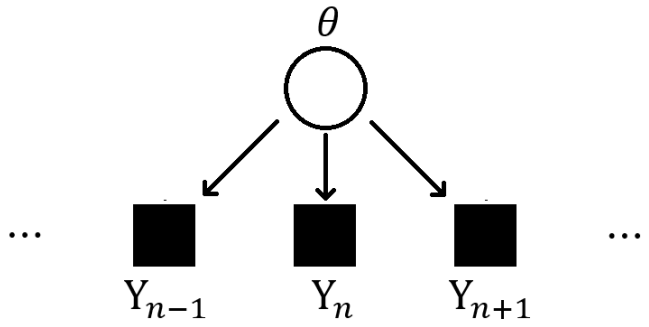


Figure A.1: A graphical representation of the typical Bayesian inference example. Here the observations Y_n's are independent and identically distributed samples controlled by the parameter θ.

and the prior $\pi(\theta)$, thus intuitively it becomes more dependent on the data as the number of data N goes to infinity. Theoretically, there are consistency theorems such as the Doob's consistency theorem and Bernstein-von Mises theorem [van der Vaart (1998)] that give good asymptotic behaviors to (A.2) under regular conditions. Figure A.1 shows the graphical representation of this example.

Now let us further consider the following example under a different setting. We still have the same objects in the above example except a set of hidden variables $\mathcal{X} = \{\mathrm{X}_n\}_{n=1}^{\mathrm{N}}$ such that the parametric model is depending them instead of the parameter: (A.3) is often referred to as the conditional independence of $\mathcal{Y}$ given $\mathcal{X}$.

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{n=1}^{\mathrm{N}} p(\mathrm{Y}_n|\mathrm{X}_n) \tag{A.3}$$

Suppose that the prior on $\mathcal{X}$ is of the naïve Bayes, i.e.,

$$p(\mathcal{X}) = \prod_{n=1}^{\mathrm{N}} p(\mathrm{X}_n) \tag{A.4}$$

Then, the posterior distribution of $\mathcal{X}$ given $\mathcal{Y}$ is given as follows:

$$p(\mathcal{X}|\mathcal{Y}) = \frac{p(\mathcal{X})\,p(\mathcal{Y}|\mathcal{X})}{p(\mathcal{Y})} \propto p(\mathcal{X})\,p(\mathcal{Y}|\mathcal{X}) = \prod_{n=1}^{\mathrm{N}} p(\mathrm{X}_n)\,p(\mathrm{Y}_n|\mathrm{X}_n) \tag{A.5}$$

(A.5) implies that each posterior of X_n is depending on only the associated observation Y_n, thus here we cannot utilize the consistency theorems and the inference is very sensitive to the choice of prior and potential outliers.

Some practical examples such as time series models give reasonable priors on $\mathcal{X}$ to utilize the dependency on X_n's that makes each posterior distribution of X_n depend on the full $\mathcal{Y}$. If such a prior is also conjugate to the model, then the posterior is derived analytically. However, in general the posterior of $\mathcal{X}$ is not given in a closed form. The next best thing would be sampling the hidden variables from their posterior or doing a maximum a posteriori (MAP) estimation. Unfortunately, even such goals are usually not tractable straightforwardly because 1) $\mathcal{X}$ could be fully correlated (fully connected in the view of the graphical representation) one another over $p(\mathcal{X}|\mathcal{Y})$ thus Markov-chain Monte Carlo algorithm on it would not make sense in a tractable number of iterations and 2) Numerical methods for maximizing $p(\mathcal{X}|\mathcal{Y})$ could not work if either $\mathcal{X}$ is defined on a finite space or $p(\mathcal{X}|\mathcal{Y})$ is not differentiable.

The Markovian assumption on the prior $p(\mathcal{X})$ looks for the balance between the naïve Bayes and fully connected assumptions. It assumes that a subset of $\mathcal{X}$ is conditionally independent of those outside of their boundary given the boundary. This memoryless property could reduce the time complexity of the algorithm drastically: details can be found in [Barber (2012)]. If the data $\mathcal{Y}$ can be arranged over $\mathbb{R}$, time for instance, then $\mathcal{X}$ might be assumed as a Markov chain, i.e., it is a sequence of events where each of them depends only on the very previous event. The graphical representation of this case can be found in figure A.2.

Generally speaking, a state-space model (SSM) [Hangos et al. (2001)] is a mathematical model of a physical system consisting of a set of input, output and state variables that are connected by differential equations. State variables evolve through time and depend on the output variables.

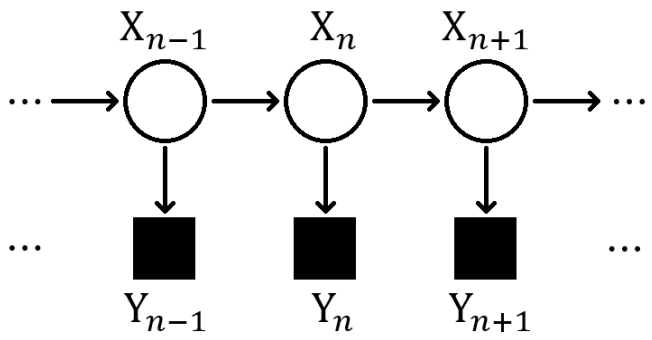


Figure A.2: A graphical representation of the typical example of the state-space model. Here the observations Y_n's are conditionally independent given the hidden states X_n's.

State-space models can be also understood as a class of probabilistic graphical model [Koller & Friedman (2009)] that describes the probabilistic dependence between the (latent) state variables and the output variables. These models are corresponding to the case where $\mathcal{X}$ is the set of latent state variables, $\mathcal{Y}$ is that of the output variables and the indices, say, $\mathcal{T} = \{T_n\}_{n=1}^N$, are input variables, in the above example.

This chapter briefly covers the probabilistic state-space models. All models in this chapter are involved with the Markovian assumption: section B.4 suggests a Gaussian process state-space model that does not take that assumption. Recurrent neural networks (RNNs) [Goodfellow et al. (2016)] are practically good choices in state-space modelling especially if data are plentiful, but this chapter does not deal with it.

A.1 Hidden Markov Model

The classification of state-space models is depending on the space in which each X_n is defined. Hidden Markov models (HMMs) [Durbin et al. (1998)] assume that each X_n is defined in a *finite* set while continuous state-space models that will be covered in section A.2 consider X_n to take continuous values. HMMs have been applied in the fields that are interested in finding latent states behind the observable sequences: gene prediction, protein folding, DNA motif discovery and chromatin state discovery are such examples in computational biology.

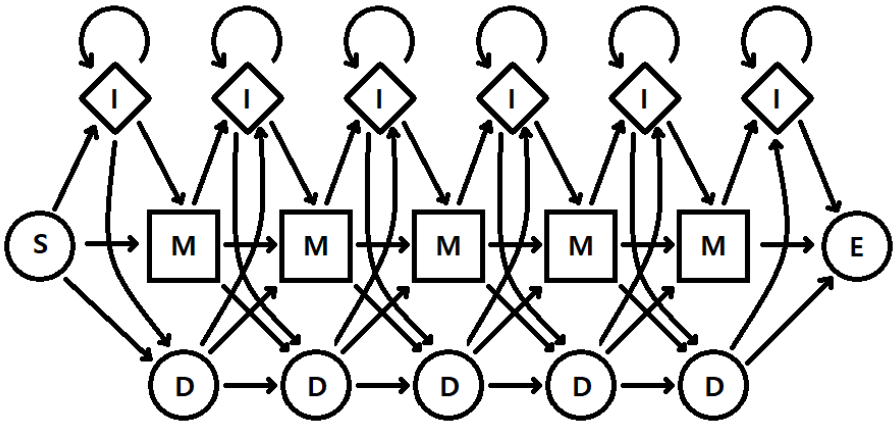


Figure A.3: Another type of the graphical representation of the hidden Markov model for the sequence alignment algorithm. Here, states S, E, I, D and M stand for start, end, insert, delete and match, respectively.

Before going further, let us first define the following notations:

- $\mathcal{Y} = \{\mathrm{Y}_n\}_{n=1}^{\mathrm{N}}$: observed sequence.
- $\mathcal{X} = \{\mathrm{X}_n\}_{n=1}^{\mathrm{N}} \subseteq \mathcal{S}$: hidden states defined in a finite set $\mathcal{S}$ that are associated with $\mathcal{Y}$.

An HMM is exactly specified by the transition and emission models. The transition model $p_t\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right)$ determines how $\mathcal{X}$ is correlated a priori and the emission model $p_e\left(\mathrm{Y}_n|\mathrm{X}_n\right)$ gives the relationship between $\mathcal{Y}$ and $\mathcal{X}$. Note that those models might be defined nonstationary, i.e., depend on the index n, but here we abuse their notations as if they were stationary. Each X_n is to take values in a finite set, so the transition model can be represented as a transition matrix and the corresponding graphical representation focuses on the transition between candidates of the hidden states rather than hidden variables themselves, like figure A.3.

The transition and emission models give the full joint likelihood of the model as follows:

$$p\left(\mathcal{X},\mathcal{Y}\right) = p\left(\mathcal{X}\right)p\left(\mathcal{Y}|\mathcal{X}\right) = \prod_{n=1}^{\mathrm{N}} p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right)\prod_{n=1}^{\mathrm{N}} p_e\left(\mathrm{Y}_n|\mathrm{X}_n\right) \tag{A.6}$$

The inference is involved with finding the MAP estimator of $\mathcal{X}$ and computing the posterior $p\left(\mathcal{X}|\mathcal{Y}\right)$: both goals can be achieved exactly by the Viterbi algorithm (section A.1.1) and Forward-backward algorithm (section A.1.2), respectively. This HMM framework can be ex-

tended to a more general setting so that the transition model has an order more than 1 (e.g., $p_t\left(\mathrm{X}_{n+1}|\mathrm{X}_n, \mathrm{X}_{n-1}\right)$) and the emission model is autoregressive (e.g., $p_e\left(\mathrm{Y}_n|\mathrm{X}_n, \mathrm{Y}_{n-1}\right)$) for each n. Here, we just focus on the model with the vanilla setting for giving the intuition. Transition and emission models are either given a priori or to be learned together with $\mathcal{X}$. Section A.1.3 covers the latter case where such models are parametric.

A.1.1 Viterbi Algorithm

The Viterbi algorithm [Viterbi (1967); Durbin et al. (1998)] is a dynamic programming for finding the MAP sequence of hidden states in the HMM given observations. Because $p\left(\mathcal{X}|\mathcal{Y}\right) \propto p\left(\mathcal{X}, \mathcal{Y}\right)$, thus maximizing $p\left(\mathcal{X}|\mathcal{Y}\right)$ is equivalent to maximizing $p\left(\mathcal{X}, \mathcal{Y}\right)$, so we have:

$$\begin{aligned}\underset{\mathcal{X}}{\operatorname{argmax}}\, p\left(\mathcal{X}|\mathcal{Y}\right) &= \underset{\mathcal{X}}{\operatorname{argmax}}\, p\left(\mathcal{X}, \mathcal{Y}\right) = \underset{\mathcal{X}}{\operatorname{argmax}}\left[\log p\left(\mathcal{X}\right) + \log p\left(\mathcal{Y}|\mathcal{X}\right)\right] \\ &= \underset{\mathcal{X}}{\operatorname{argmax}}\left[\sum_{n=1}^{\mathrm{N}} \log p_e\left(\mathrm{Y}_n|\mathrm{X}_n\right) + \sum_{n=1}^{\mathrm{N}} \log p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right)\right]\end{aligned} \tag{A.7}$$

Thus, finding the MAP sequence of $p\left(\mathcal{X}|\mathcal{Y}\right)$ is again equivalent to obtaining the most probable path of $\mathcal{L}$ in (A.8):

$$\mathcal{L}\left(\mathcal{X}\right) \triangleq \sum_{n=1}^{\mathrm{N}} \log p_e\left(\mathrm{Y}_n|\mathrm{X}_n\right) + \sum_{n=1}^{\mathrm{N}} \log p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right) \tag{A.8}$$

Let us define a recursive term $v_n\left(s\right)$ for each $s \in \mathcal{S}$ and $n = 2, 3, \cdots, \mathrm{N}$ as follows:

$$\begin{aligned}v_1\left(s\right) &\triangleq \log p_e\left(\mathrm{Y}_1|\mathrm{X}_1 = s\right) + \log p\left(\mathrm{X}_1 = s\right) \\ v_n\left(s\right) &\triangleq \log p_e\left(\mathrm{Y}_n|\mathrm{X}_n = s\right) + \max_{k \in \mathcal{S}}\left(v_{n-1}\left(k\right) + \log p_t\left(\mathrm{X}_n = s|\mathrm{X}_{n-1} = k\right)\right)\end{aligned} \tag{A.9}$$

Then, one can observe that the following holds for each $m = 1, 2, \cdots, \mathrm{N}-1$:

$$\begin{aligned}&\max_{\mathrm{X}_{m:\mathrm{N}}} \sum_{n=m+1}^{\mathrm{N}} \log p_e\left(\mathrm{Y}_n|\mathrm{X}_n\right) + \sum_{n=m+1}^{\mathrm{N}} \log p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right) + v_m\left(\mathrm{X}_m\right) \\ &= \max_{\mathrm{X}_{m+1:\mathrm{N}}} \sum_{n=m+2}^{\mathrm{N}} \log p_e\left(\mathrm{Y}_n|\mathrm{X}_n\right) + \sum_{n=m+2}^{\mathrm{N}} \log p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right) + v_{m+1}\left(\mathrm{X}_{m+1}\right)\end{aligned} \tag{A.10}$$

Thus, $\max_{\mathcal{X}} \mathcal{L}\left(\mathcal{X}\right) = \max_{\mathrm{X}_\mathrm{N}} v_\mathrm{N}\left(\mathrm{X}_\mathrm{N}\right)$. Note that computing each $v_n\left(s\right)$ requires the time complexity $\mathcal{O}\left(|\mathcal{S}|\right)$, so the time complexity of computing $\max_{\mathcal{X}} \mathcal{L}\left(\mathcal{X}\right)$ is of $\mathcal{O}\left(|\mathcal{S}|^2 \mathrm{N}\right)$ and the algorithm requires the storage of $\mathcal{O}\left(|\mathcal{S}| \mathrm{N}\right)$.

The goal is to find the maximizer of $\mathcal{L}\left(\mathcal{X}\right)$, i.e., $\hat{\mathcal{X}}$ such that $\mathcal{L}\left(\hat{\mathcal{X}}\right) = \max_{\mathcal{X}} \mathcal{L}\left(\mathcal{X}\right)$. Let us define another recursive term $b_n\left(s\right)$ for each $s \in \mathcal{S}$ and $n = 1, 2, \cdots, \mathrm{N}-1$ as follows:

$$b_n\left(s\right) \triangleq \operatorname*{argmax}_{k \in \mathcal{S}} \left(v_n\left(k\right) + \log p_t\left(\mathrm{X}_{n+1} = s|\mathrm{X}_n = k\right)\right) \tag{A.11}$$

Note that $\hat{\mathrm{X}}_\mathrm{N} = \operatorname{argmax}_{\mathrm{X}_\mathrm{N}} v_\mathrm{N}\left(\mathrm{X}_\mathrm{N}\right)$ and $\hat{\mathrm{X}}_{\mathrm{N}-1} = b_{\mathrm{N}-1}\left(\hat{\mathrm{X}}_\mathrm{N}\right)$ because:

$$\begin{aligned}&\max_{\mathcal{X}} \mathcal{L}\left(\mathcal{X}\right) = \max_{\mathrm{X}_\mathrm{N}} v_\mathrm{N}\left(\mathrm{X}_\mathrm{N}\right) \\ &= \log p_e\left(\mathrm{Y}_\mathrm{N}\middle|\hat{\mathrm{X}}_\mathrm{N}\right) + \max_{k \in \mathcal{S}} \left(v_{\mathrm{N}-1}\left(k\right) + \log p_t\left(\mathrm{X}_\mathrm{N} = \hat{\mathrm{X}}_\mathrm{N}\middle|\mathrm{X}_{\mathrm{N}-1} = k\right)\right)\end{aligned} \tag{A.12}$$

Moreover, one can show that the following holds for each $m = 1, 2, \cdots, \mathrm{N}-2$:

$$\begin{aligned}\mathcal{L}\left(\hat{\mathcal{X}}\right) = &\sum_{n=m+1}^{\mathrm{N}} \log p_e\left(\mathrm{Y}_n\middle|\hat{\mathrm{X}}_n\right) + \sum_{n=m+1}^{\mathrm{N}-1} \log p_e\left(\hat{\mathrm{X}}_{n+1}\middle|\hat{\mathrm{X}}_n\right) \\ &+ \max_{k \in \mathcal{S}} \left(v_m\left(k\right) + \log p_t\left(\mathrm{X}_{m+1} = \hat{\mathrm{X}}_{m+1}\middle|\mathrm{X}_m = k\right)\right)\end{aligned} \tag{A.13}$$

Thus, $\hat{\mathrm{X}}_n = b_n\left(\hat{\mathrm{X}}_{n+1}\right)$ holds for all $n = 1, 2, \cdots, \mathrm{N}-1$. Since each $b_n\left(s\right)$ can be stored in the step for updating v_n's with time complexity $\mathcal{O}\left(|\mathcal{S}|\right)$, the time complexity of computing $\operatorname{argmax}_{\mathcal{X}} \mathcal{L}\left(\mathcal{X}\right) = \operatorname{argmax}_{\mathcal{X}} p\left(\mathcal{X}|\mathcal{Y}\right)$ is still $\mathcal{O}\left(|\mathcal{S}|^2 \mathrm{N}\right)$.

So far, we have observed how the MAP estimator of the hidden states given observations is

obtained by the Viterbi algorithm with time complexity $\mathcal{O}\left(|\mathcal{S}|^2 \mathrm{N}\right)$, i.e., linear to the length of a sequence and quadratic to the size of a hidden state space. Though this algorithm is still popular, it has the following drawbacks:

1. This deterministic algorithm only returns the most probable path, thus is not proper if multiple probable paths are expected and to be considered one-by-one.

2. It assumes that both transition and emission models are given a priori, which is not common in practice.

The above downside will be addressed in the following subsections.

A.1.2 Forward-backward Algorithm

The forward-backward algorithm [Durbin et al. (1998)] is a probabilistic approach to HMMs. It focuses on computing the posterior distribution of hidden states and drawing sample paths from the posterior. As shown in its name, the forward-backward algorithm consists of the forward and backward algorithm. The forward algorithm recursively computes the forward sum $p\left(\mathrm{Y}_{1:n}, \mathrm{X}_n = s\right)$ while the backward algorithm updates $p\left(\mathrm{Y}_{n+1:\mathrm{N}}|\mathrm{X}_n = s\right)$ for each n and $s \in \mathcal{S}$.

Note that the following holds for any $1 \leq l \leq m \leq \mathrm{N}$: let $p\left(\mathrm{Y}_{\mathrm{N}+1:\mathrm{N}}|\mathrm{X}_m\right) \equiv 1$ in notation for saving the space.

$$\begin{aligned} p\left(\mathrm{X}_{l:m}|\mathrm{Y}_{1:\mathrm{N}}\right) &\propto p\left(\mathrm{X}_{l:m}, \mathrm{Y}_{1:\mathrm{N}}\right) \\ &= p\left(\mathrm{Y}_{m+1:\mathrm{N}}|\mathrm{X}_m\right) p\left(\mathrm{Y}_{1:l}, \mathrm{X}_l\right) \prod_{n=l+1}^{m} p\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right) p\left(\mathrm{Y}_n|\mathrm{X}_n\right) \end{aligned} \tag{A.14}$$

, and, as a special case, $p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right) \propto p\left(\mathrm{Y}_{n+1:\mathrm{N}}|\mathrm{X}_n\right) p\left(\mathrm{Y}_{1:n}, \mathrm{X}_n\right)$.

For the forward algorithm, the following holds for each $n = 2, 3, \cdots, \mathrm{N}$ to run the algorithm

recursively:

$$\begin{aligned} p\left(\mathrm{Y}_1, \mathrm{X}_1 = s\right) &= p\left(\mathrm{Y}_1 | \mathrm{X}_1 = s\right) p\left(\mathrm{X}_1 = s\right) \\ p\left(\mathrm{Y}_{1:n}, \mathrm{X}_n = s\right) &= p\left(\mathrm{Y}_n | \mathrm{X}_n = s\right) \sum_{k \in \mathcal{S}} p\left(\mathrm{X}_n = s | \mathrm{X}_{n-1} = k\right) p\left(\mathrm{Y}_{1:n-1}, \mathrm{X}_{n-1} = k\right) \end{aligned} \tag{A.15}$$

One can compute the marginal likelihood $p\left(\mathrm{Y}_{1:\mathrm{N}}\right) = \sum_{s \in \mathcal{S}} p\left(\mathrm{Y}_{1:\mathrm{N}}, \mathrm{X}_{\mathrm{N}} = s\right)$ from the last forward term.

Similarly, the backward algorithm is iterated based on the following equality for each $n = \mathrm{N}-1, \mathrm{N}-2, \cdots, 1$:

$$p\left(\mathrm{Y}_{n+1:\mathrm{N}} | \mathrm{X}_n = s\right) = \sum_{k \in \mathcal{S}} p\left(\mathrm{X}_{n+1} = k | \mathrm{X}_n = s\right) p\left(\mathrm{Y}_{n+1} | \mathrm{X}_{n+1} = k\right) p\left(\mathrm{Y}_{n+2:\mathrm{N}} | \mathrm{X}_{n+1} = k\right) \tag{A.16}$$

Note that computing each forward/backward term has time complexity of $\mathcal{O}\left(|\mathcal{S}|\right)$, thus that of the algorithm is $\mathcal{O}\left(|\mathcal{S}|^2 \mathrm{N}\right)$ in time and $\mathcal{O}\left(|\mathcal{S}| \mathrm{N}\right)$ in storage. Thus, the forward-backward algorithm gives a way to compute the posterior of each hidden state efficiently.

The backward sampling algorithm is a variant of the forward-backward algorithm. Instead of computing the backward terms $p\left(\mathrm{Y}_{n+1:\mathrm{N}} | \mathrm{X}_n = s\right)$'s, the algorithm samples $\mathcal{X}$ from the posterior $p\left(\mathcal{X} | \mathcal{Y}\right)$ backwards by utilizing the forward terms with the following relationships:

$$\begin{aligned} p\left(\mathrm{X}_{\mathrm{N}} = s | \mathrm{Y}_{1:\mathrm{N}}\right) &\propto p\left(\mathrm{Y}_{1:\mathrm{N}}, \mathrm{X}_{\mathrm{N}} = s\right) \\ p\left(\mathrm{X}_n = s \middle| \mathrm{X}_{n+1} = \widetilde{\mathrm{X}}_{n+1}, \mathrm{Y}_{1:\mathrm{N}}\right) &\propto p\left(\mathrm{X}_{n+1} = \widetilde{\mathrm{X}}_{n+1} \middle| \mathrm{X}_n = s\right) p\left(\mathrm{Y}_{1:n}, \mathrm{X}_n = s\right) \end{aligned} \tag{A.17}$$

Note that the backward sampling in (A.17) is of time complexity $\mathcal{O}\left(|\mathcal{S}| \mathrm{N}\right)$, but the full algorithm is still $\mathcal{O}\left(|\mathcal{S}|^2 \mathrm{N}\right)$ because it uses the forward terms that are obtained in the forward algorithm.

In practice, the algorithm is usually challenged by the underflow, i.e., the forward and backward terms are too small not to be regarded as 0 in computer if the length of sequence N is large. One solution is to store the logarithmic values of the forward and backward terms instead of the

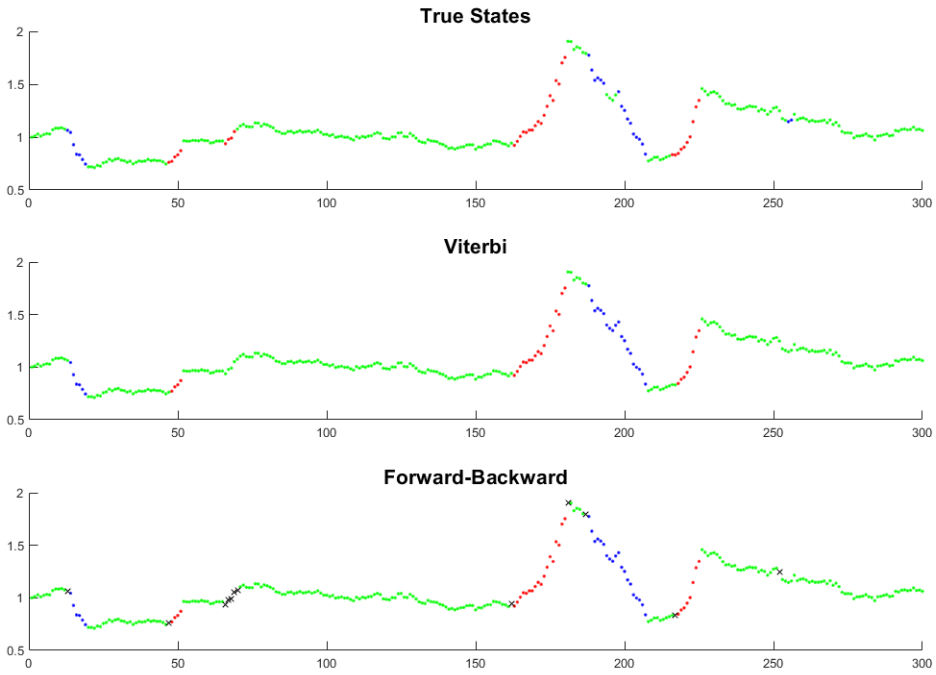


Figure A.4: An example of HMM and the inference by the Viterbi and forward-backward algorithm.

original values. It utilizes the following log-sum-exp trick:

$$\log \sum_{n=1}^{\mathrm{N}} \exp\left(x_n\right) = \max_m x_m + \log \sum_{n=1}^{\mathrm{N}} \exp\left(x_n - \max_m x_m\right) \tag{A.18}$$

Figure A.4 shows an example of the HMM for the time series analysis. The upper panel shows the true hidden states in colors: blue dots are the observations suffering from the crash, red dots are those going on soaring, and green dots move sideways. The middle and bottom panels show the inferred hidden states by the Viterbi and forward-backward algorithms, respectively. One advantage of the forward-backward algorithm is that the inference is probabilistic so does more careful analysis: black crosses indicate that the associated MAP probabilities are less than 0.7, i.e., less certain than the others.

A.1.3 Baum-Welch Algorithm

So far, we have assumed that the transition and emission models are given. However, in practice both transition and emission models are not given a priori. Here those models are assumed to be parametric.

The Baum-Welch algorithm [Durbin et al. (1998)] is an expectation-maximization (EM) algorithm [Dempster et al. (1977)] that infers unknown parameters of HMMs. Let us formulate the full joint likelihood of $\mathcal{X}$ and $\mathcal{Y}$, where transition and emission models are parametrized by θ and ψ, respectively.

$$p\left(\mathcal{X}, \mathcal{Y}, \theta, \psi\right) = \pi\left(\theta, \psi\right) p\left(\mathcal{X}, \mathcal{Y} | \theta, \psi\right) = \pi\left(\theta, \psi\right) \prod_{n=1}^{\mathrm{N}} p_t\left(\mathrm{X}_n | \mathrm{X}_{n-1}, \theta\right) \prod_{n=1}^{\mathrm{N}} p_e\left(\mathrm{Y}_n | \mathrm{X}_n, \psi\right) \tag{A.19}$$

, where $\pi\left(\theta, \psi\right)$ is the prior on the parameters.

The EM algorithm is consisting of the expectation step (E-step) and maximization step (M-step). Here $\mathcal{X}$ is the hidden state, so the E-step computes the following expectation for the previously updated parameters $\theta^{(t)}$, $\psi^{(t)}$:

$$\begin{aligned}
&\mathcal{Q}\left(\theta, \psi \middle| \theta^{(t)}, \psi^{(t)}\right) \triangleq \mathbb{E}_{\mathcal{X}|\mathcal{Y}, \theta^{(t)}, \psi^{(t)}}\left[\log p\left(\mathcal{X}, \mathcal{Y}, \theta, \psi\right)\right] \\
&= \mathbb{E}_{\mathcal{X}|\mathcal{Y}, \theta^{(t)}, \psi^{(t)}}\left[\log \pi\left(\theta, \psi\right) + \sum_{n=1}^{\mathrm{N}} \log p_t\left(\mathrm{X}_n | \mathrm{X}_{n-1}, \theta\right) + \sum_{n=1}^{\mathrm{N}} \log p_e\left(\mathrm{Y}_n | \mathrm{X}_n, \psi\right)\right] \\
&= \log \pi\left(\theta, \psi\right) + \sum_{n=1}^{\mathrm{N}} \sum_{\mathrm{X}_n \in \mathcal{S}} \sum_{\mathrm{X}_{n-1} \in \mathcal{S}} \mathbb{P}\left(\mathrm{X}_{n-1}, \mathrm{X}_n\right) \log p_t\left(\mathrm{X}_n | \mathrm{X}_{n-1}, \theta\right) \\
&+ \sum_{n=1}^{\mathrm{N}} \sum_{\mathrm{X}_n \in \mathcal{S}} \mathbb{P}\left(\mathrm{X}_n\right) \log p_e\left(\mathrm{Y}_n | \mathrm{X}_n, \psi\right)
\end{aligned} \tag{A.20}$$

, where $\mathbb{P}\left(\mathrm{X}_n\right) \triangleq p\left(\mathrm{X}_n \middle| \mathrm{Y}_{1:\mathrm{N}}, \theta^{(t)}, \psi^{(t)}\right)$ and $\mathbb{P}\left(\mathrm{X}_{n-1}, \mathrm{X}_n\right) \triangleq p\left(\mathrm{X}_{n-1:n} \middle| \mathrm{Y}_{1:\mathrm{N}}, \theta^{(t)}, \psi^{(t)}\right)$ for each n. Note that $\mathbb{P}$ is given by the forward-backward algorithm in section A.1.2.

The M-step finds the parameters that maximize $\mathcal{Q}\left(\theta, \psi \middle| \theta^{(t)}, \psi^{(t)}\right)$, i.e.,

$$\left(\theta^{(t+1)}, \psi^{(t+1)}\right) = \underset{\theta,\psi}{\operatorname{argmax}}\, \mathcal{Q}\left(\theta, \psi \middle| \theta^{(t)}, \psi^{(t)}\right) \tag{A.21}$$

Both models and hidden states can be inferred by iterating the E-step that computes $\mathbb{P}$ from the forward-backward algorithm and the M-step that optimizes (A.21) until convergence. Note that, however, the EM algorithm only guarantees the local optimum, thus limited in such a sense.

The optimization of (A.21) can be done by either numerically (e.g., gradient descent) or analytically, which depends on how transition and emission models are parametrized. For example, suppose that $\pi \equiv 1$, the observations $\mathcal{Y}$ are also defined in a finite set $\mathcal{V}$, both transition and emission models are stationary and θ and ψ indicate transition and emission tables, respectively. Then, one can derive the following updating formulations analytically:

$$\begin{aligned} \theta^{(t+1)}(s,k) &= \frac{\sum_{n=1}^{\mathrm{N}} \mathbb{P}\left(\mathrm{X}_{n-1} = s, \mathrm{X}_n = k\right)}{\sum_{n=1}^{\mathrm{N}} \mathbb{P}\left(\mathrm{X}_{n-1} = s\right)} \\ \psi^{(t+1)}(s,v) &= \frac{\sum_{n=1}^{\mathrm{N}} \mathbf{1}_{\mathrm{Y}_n = v} \cdot \mathbb{P}\left(\mathrm{X}_n = s\right)}{\sum_{n=1}^{\mathrm{N}} \mathbb{P}\left(\mathrm{X}_n = s\right)} \end{aligned} \tag{A.22}$$

, where $\theta(s,k) \triangleq p_t\left(\mathrm{X}_n = k | \mathrm{X}_{n-1} = s\right)$ and $\psi(s,v) \triangleq p_e\left(\mathrm{Y}_n = v | \mathrm{X}_n = s\right)$. The Baum-Welch algorithm infers both models and hidden states by iterating the E-step that computes $\mathbb{P}$ from the forward-backward algorithm and the M-step that updates the transition and emission models by (A.22) until convergence.

A.1.4 Variational Approximation

As discussed earlier, the exact inference of a fully connected hidden state model is usually intractable. One detour is to approximate it with a tractable Markovian model. This subsection briefly introduces the core idea regarding that methodology.

Let us assume that the real joint distribution of $(\mathcal{X}, \mathcal{Y})$ is $p(\mathcal{X}, \mathcal{Y}) = p(\mathcal{Y}|\mathcal{X})\, p(\mathcal{X})$ and let $q(\mathcal{X}) = q(\mathcal{X}|\theta)$ be a tractable Markov chain parametrized by θ for approximating $p(\mathcal{X}|\mathcal{Y})$.

Then, because the Kullback-Leibler (KL) divergence is always nonnegative, we have the following:

$$\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{Y}\right)\right) = \log p\left(\mathcal{Y}\right) + \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\right) - \mathbb{E}_{q(\mathcal{X}|\theta)}\left[\log p\left(\mathcal{Y}|\mathcal{X}\right)\right] \geq 0 \tag{A.23}$$

Therefore, the log-marginal likelihood $\log p\left(\mathcal{Y}\right)$ has the evidence lower bound (ELBO) $\mathcal{L}$ defined as follows:

$$\log p\left(\mathcal{Y}\right) \geq \mathbb{E}_{q(\mathcal{X}|\theta)}\left[\log p\left(\mathcal{Y}|\mathcal{X}\right)\right] - \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\right) = \mathcal{L} \tag{A.24}$$

Now the goal is to maximize $\mathcal{L}$ instead of $\log p\left(\mathcal{Y}\right)$ directly: this is the core idea of variational approximation [Bishop (2006)]. Note that $\log p\left(\mathcal{Y}|\mathcal{X}\right)$ is easily decomposed into the summation if $\mathcal{Y}$ is assumed to be conditionally independent given $\mathcal{X}$ and $\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\right)$ can be expressed analytically as a summation in discrete modelling.

Once q is learned, one can utilize it for the sampling purpose. Note that the sample paths drawn from the variational approximation q is not the exact ones but can be used for the initialization in the Markov-chain Monte Carlo (MCMC) algorithm that will be covered in section A.4. If obtaining posterior distributions of each X_n is interesting only, defining q as a naïve Bayes could be an alternative: it directly approximates the posteriors then.

A.2 Kalman Filter and Smoother

In section A.1, we discussed the case that hidden states are defined in a finite set. Briefly speaking, Viterbi and forward-backward algorithms for HMMs give ways of considering all possible assignments to the hidden states efficiently. However, this cannot be extended to the case that hidden states are defined continuously: there are infinitely many possible assignments. Nonetheless, computing the analogues of forward and backward terms in forms of continuous distributions is still feasible under some conditions. This section covers Kalman filter (section A.2.1) and smoother (section A.2.2) for the cases that both transition and emission models are Gaussian.

In practice, Gaussian assumption could be too strong to model the data: mixture Kalman filter (section A.2.3) would be an alternative to alleviate such strong conditions.

A.2.1 Kalman Filter

The Kalman filter [Kalman (1960); Murphy (2012)] is an algorithm that uses a series of noisy observations to estimate the underlying hidden states that are associated to them. Similar to HMMs, it consists of the transition and emission models: sometimes the model also assumes the control-input model that intentionally affect the hidden states in each step, but it is not considered in this dissertation. The following notations are assumed in the rest of section A.2.

- $\mathcal{X} = \{\mathrm{X}_n\}_{n=0}^{\mathrm{N}}$: a series of hidden states where each $\mathrm{X}_n \in \mathbb{R}^{\mathrm{D}}$ and X_0 is an (hypothetical) initial state.
- $\mathcal{Y} = \{\mathrm{Y}_n\}_{n=1}^{\mathrm{N}}$: a series of observations where each $\mathrm{Y}_n \in \mathbb{R}^{\mathrm{L}}$.
- $\mathcal{F} = \{\mathrm{F}_n\}_{n=1}^{\mathrm{N}}$: a set of transition models where a $\mathrm{D} \times \mathrm{D}$ matrix F_n affects the transition from X_{n-1} to X_n.
- $\mathcal{H} = \{\mathrm{H}_n\}_{n=1}^{\mathrm{N}}$: a set of emission models where a $\mathrm{L} \times \mathrm{D}$ matrix H_n affects the emission from X_n to Y_n.
- $\mathcal{Q} = \{\mathrm{Q}_n\}_{n=1}^{\mathrm{N}}$: a set of covariance matrices in transition models where a $\mathrm{D} \times \mathrm{D}$ matrix Q_n is involved with the transition from X_{n-1} to X_n.
- $\mathcal{B} = \{b_n\}_{n=1}^{\mathrm{N}}$: a set of bias vectors in transition models where each $b_n \in \mathbb{R}^{\mathrm{D}}$ affects the transition from X_{n-1} to X_n.
- $\mathcal{R} = \{\mathrm{R}_n\}_{n=1}^{\mathrm{N}}$: a set of covariance matrices in emission models where a $\mathrm{L} \times \mathrm{L}$ matrix R_n is involved with the emission from X_n to Y_n.
- $\mathcal{C} = \{c_n\}_{n=1}^{\mathrm{N}}$: a set of bias vectors in emission models where each $c_n \in \mathbb{R}^{\mathrm{L}}$ affects the emission from X_n to Y_n.

The transition $p_t = \left\{p_t^{(n)}\right\}_{n=1}^{\mathrm{N}}$ and emission $p_e = \left\{p_e^{(n)}\right\}_{n=1}^{\mathrm{N}}$ models are given as follows:

$$\begin{aligned} p_t^{(n)}\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right) &= \mathcal{N}\left(\mathrm{X}_n|\mathrm{F}_n\mathrm{X}_{n-1}+b_n,\mathrm{Q}_n\right) \\ p_e^{(n)}\left(\mathrm{Y}_n|\mathrm{X}_n\right) &= \mathcal{N}\left(\mathrm{Y}_n|\mathrm{H}_n\mathrm{X}_n+c_n,\mathrm{R}_n\right) \end{aligned} \tag{A.25}$$

The Kalman filter aims at computing the prediction term $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n-1}\right)$, forward posterior $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ and posterior predictive $p\left(\mathrm{Y}_n|\mathrm{Y}_{1:n-1}\right)$ for each n: this approach is more realistic if the data are being received on real time and one is supposed to make predictions promptly. Suppose that $p\left(\mathrm{X}_{n-1}|\mathrm{Y}_{1:n-1}\right) = \mathcal{N}\left(\mathrm{X}_{n-1}\middle|\mu_{n-1},\Sigma_{n-1}\right)$. Then, one can derive the following:

$$\begin{aligned} p\left(\mathrm{X}_n|\mathrm{Y}_{1:n-1}\right) &= \mathcal{N}\left(\mathrm{X}_n\middle|\mu_{n|n-1},\Sigma_{n|n-1}\right) \\ &\triangleq \mathcal{N}\left(\mathrm{X}_n\middle|\mathrm{F}_n\mu_{n-1}+b_n,\mathrm{Q}_n+\mathrm{F}_n\Sigma_{n-1}\mathrm{F}_n^{\mathbb{T}}\right) \\ &= \int p_t^{(n)}\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right)p\left(\mathrm{X}_{n-1}|\mathrm{Y}_{1:n-1}\right)d\mathrm{X}_{n-1} \end{aligned} \tag{A.26}$$

$$\begin{aligned} p\left(\mathrm{Y}_n|\mathrm{Y}_{1:n-1}\right) &= \mathcal{N}\left(\mathrm{X}_n\middle|\hat{\mu}_n,\hat{\Sigma}_n\right) \\ &\triangleq \mathcal{N}\left(\mathrm{Y}_n\middle|\mathrm{H}_n\mu_{n|n-1}+c_n,\mathrm{R}_n+\mathrm{H}_n\Sigma_{n|n-1}\mathrm{H}_n^{\mathbb{T}}\right) \\ &= \int p_e^{(n)}\left(\mathrm{Y}_n|\mathrm{X}_n\right)p\left(\mathrm{X}_n|\mathrm{Y}_{1:n-1}\right)d\mathrm{X}_n \end{aligned} \tag{A.27}$$

$$\begin{aligned} &p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) = \mathcal{N}\left(\mathrm{X}_n\middle|\mu_n,\Sigma_n\right) \\ &\triangleq \mathcal{N}\left(\mathrm{X}_n\middle|\mu_{n|n-1}+\Sigma_{n|n-1}\mathrm{H}_n^{\mathbb{T}}\hat{\Sigma}_n^{-1}\left(\mathrm{Y}_n-\hat{\mu}_n\right),\Sigma_{n|n-1}-\Sigma_{n|n-1}\mathrm{H}_n^{\mathbb{T}}\hat{\Sigma}_n^{-1}\mathrm{H}_n\Sigma_{n|n-1}\right) \\ &\propto p_e^{(n)}\left(\mathrm{Y}_n|\mathrm{X}_n\right)p\left(\mathrm{X}_n|\mathrm{Y}_{1:n-1}\right) \end{aligned} \tag{A.28}$$

The terms to compute can be recursively updated by iterating (A.26), (A.27) and (A.28). The initial term $p\left(\mathrm{X}_1|\mathrm{Y}_1\right) = \mathcal{N}\left(\mathrm{X}_1\middle|\mu_1,\Sigma_1\right) \propto p_e^{(1)}\left(\mathrm{Y}_1|\mathrm{X}_1\right)p_t^{(1)}\left(\mathrm{X}_1|\mathrm{X}_0\right)$ is given as follows:

$$\begin{aligned} \mu_1 &= \mathrm{F}_1\mathrm{X}_0+b_1+\mathrm{Q}_1\mathrm{H}_1^{\mathbb{T}}\left(\mathrm{R}_1+\mathrm{H}_1\mathrm{Q}_1\mathrm{H}_1^{\mathbb{T}}\right)^{-1}\left(\mathrm{Y}_1-\mathrm{H}_1\left(\mathrm{F}_1\mathrm{X}_0+b_1\right)-c_1\right) \\ \Sigma_1 &= \mathrm{Q}_1-\mathrm{Q}_1\mathrm{H}_1^{\mathbb{T}}\left(\mathrm{R}_1+\mathrm{H}_1\mathrm{Q}_1\mathrm{H}_1^{\mathbb{T}}\right)^{-1}\mathrm{H}_1\mathrm{Q}_1 \end{aligned} \tag{A.29}$$

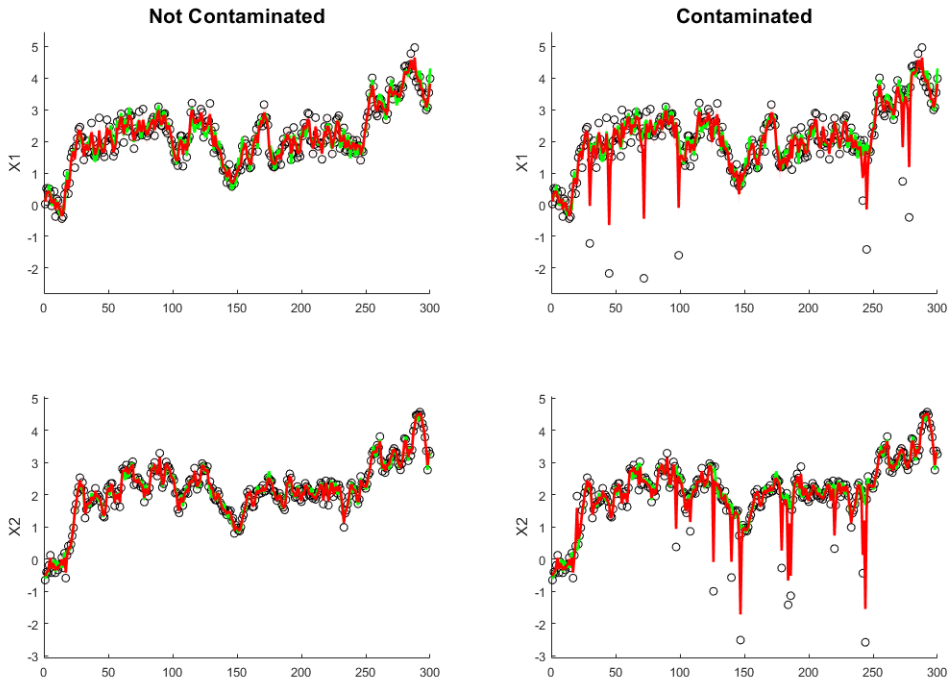


Figure A.5: An example of Kalman smoother where the 2-dimensional noisy observations are filtered. For each panel, black circles are the observations, the green curve is the true denoised observations (hidden states) and red curve is the inferred hidden states from the noisy observations. The left and right panels use the same observations except for 5% of the contamination in the right panels.

Note that the terms in (A.26), (A.27), (A.28) and (A.29) are expressed in closed forms because their integrations and proportionality are involved with Gaussian terms, thus not extendable to the cases with general transition and observation models.

A.2.2 Kalman Smoother

The Kalman smoother [Rauch et al. (1965)] is an analogue of the forward-backward algorithm for HMMs, i.e., it computes the posterior distribution $p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right)$ for each n *after* observing all $\mathrm{Y}_{1:\mathrm{N}}$. There are two derivations: one updates $p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right)$ from $p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:\mathrm{N}}\right)$ one-by-one backwards [Murphy (2012)] and the other is two-filter smoothing [Kitagawa (1994); Briers et al. (2010)] that recursively computes $p\left(\mathrm{X}_n|\mathrm{Y}_{n+1:\mathrm{N}}\right)$ backwards and restores each $p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right)$ from $p\left(\mathrm{X}_n|\mathrm{Y}_{n+1:\mathrm{N}}\right)$ and $p\left(\mathrm{X}_n|\mathrm{Y}_{n:\mathrm{N}}\right)$, just as the forward-backward algorithm. This section follows the former one.

From (A.25) to (A.28), we have the following:

$$\begin{aligned} p\left(\mathrm{X}_n, \mathrm{X}_{n+1} | \mathrm{Y}_{1:n}\right) &= p_t^{(n)}\left(\mathrm{X}_{n+1} | \mathrm{X}_n\right) p\left(\mathrm{X}_n | \mathrm{Y}_{1:n}\right) \\ &= \mathcal{N}\left(\mathrm{X}_{n+1} | \mathrm{F}_{n+1}\mathrm{X}_n + b_{n+1}, \mathrm{Q}_{n+1}\right) \mathcal{N}\left(\mathrm{X}_n \middle| \mu_n, \Sigma_n\right) \\ &= \mathcal{N}\left(\begin{pmatrix} \mathrm{X}_n \\ \mathrm{X}_{n+1} \end{pmatrix} \middle| \begin{pmatrix} \mu_n \\ \mu_{n+1|n} \end{pmatrix}, \begin{pmatrix} \Sigma_n & \Sigma_n \mathrm{F}_{n+1}^{\mathbb{T}} \\ \mathrm{F}_{n+1}\Sigma_n & \Sigma_{n+1|n} \end{pmatrix}\right) \end{aligned} \tag{A.30}$$

, thus:

$$\begin{aligned} &p\left(\mathrm{X}_n | \mathrm{X}_{n+1}, \mathrm{Y}_{1:n}\right) \\ &= \mathcal{N}\left(\mathrm{X}_n \middle| \mu_n + \Sigma_n \mathrm{F}_{n+1}^{\mathbb{T}} \Sigma_{n+1|n}^{-1}\left(\mathrm{X}_{n+1} - \mu_{n+1|n}\right), \Sigma_n - \Sigma_n \mathrm{F}_{n+1}^{\mathbb{T}} \Sigma_{n+1|n}^{-1} \mathrm{F}_{n+1} \Sigma_n\right) \end{aligned} \tag{A.31}$$

Suppose that $p\left(\mathrm{X}_{n+1} | \mathrm{Y}_{1:\mathrm{N}}\right) = \mathcal{N}\left(\mathrm{X}_{n+1} \middle| \mu_{n+1|\mathrm{N}}, \Sigma_{n+1|\mathrm{N}}\right)$ is given. Then, one can derive the following with (A.31):

$$\begin{aligned} p\left(\mathrm{X}_n | \mathrm{Y}_{1:\mathrm{N}}\right) &= \mathcal{N}\left(\mathrm{X}_n \middle| \mu_{n|\mathrm{N}}, \Sigma_{n|\mathrm{N}}\right) \\ &\triangleq \int p\left(\mathrm{X}_n | \mathrm{X}_{n+1}, \mathrm{Y}_{1:n}\right) p\left(\mathrm{X}_{n+1} | \mathrm{Y}_{1:\mathrm{N}}\right) d\mathrm{X}_{n+1} \end{aligned} \tag{A.32}$$

, where:

$$\begin{aligned} \mu_{n|\mathrm{N}} &= \mu_n + \Sigma_n \mathrm{F}_{n+1}^{\mathbb{T}} \Sigma_{n+1|n}^{-1}\left(\mu_{n+1|\mathrm{N}} - \mu_{n+1|n}\right) \\ \Sigma_{n|\mathrm{N}} &= \Sigma_n + \Sigma_n \mathrm{F}_{n+1}^{\mathbb{T}} \Sigma_{n+1|n}^{-1}\left(\Sigma_{n+1|\mathrm{N}} - \Sigma_{n+1|n}\right) \Sigma_{n+1|n}^{-1} \mathrm{F}_{n+1} \Sigma_n \end{aligned} \tag{A.33}$$

In addition, the initial term $p\left(\mathrm{X}_{\mathrm{N}} | \mathrm{Y}_{1:\mathrm{N}}\right) = \mathrm{N}\left(\mathrm{X}_{\mathrm{N}} \middle| \mu_{\mathrm{N}|\mathrm{N}}, \Sigma_{\mathrm{N}|\mathrm{N}}\right)$ is given directly from (A.28). To sum up, the posterior distribution of each X_n given the full observation $\mathrm{Y}_{1:\mathrm{N}}$ is again a Gaussian distribution with explicit mean vector and covariance matrix that can be computed backwards recursively.

Since $p\left(\mathrm{X}_n | \mathrm{X}_{n+1}, \mathrm{Y}_{1:\mathrm{N}}\right) = p\left(\mathrm{X}_n | \mathrm{X}_{n+1}, \mathrm{Y}_{1:n}\right)$, backward sampling is easily done by (A.31): $\widetilde{\mathrm{X}}_{\mathrm{N}}$ is drawn from $p\left(\mathrm{X}_{\mathrm{N}} | \mathrm{Y}_{1:\mathrm{N}}\right) = \mathcal{N}\left(\mathrm{X}_{\mathrm{N}} \middle| \mu_{\mathrm{N}|\mathrm{N}}, \Sigma_{\mathrm{N}|\mathrm{N}}\right)$ and $\widetilde{\mathrm{X}}_n$ from $p\left(\mathrm{X}_n \middle| \widetilde{\mathrm{X}}_{n+1}, \mathrm{Y}_{1:n}\right)$ backwards one-by-one recursively.

A.2.3 Mixture Kalman Filter

So far, we have derived the Kalman filter and smoother that assume Gaussian transition and emission models. In reality, it is natural that either or both of transition and emission models are non-Gaussian and figure A.5 shows that even a small portion of contamination can ruin the inference by Kalman smoother. There have been a lot of approximation methods for such cases, and here the mixture Kalman filter [Chen & Liu (2000)], an example of the Rao-Blackwellized particle filters that reduce the sample variance much, is discussed as one of them.

First of all, let us redefine the notations:

- $\mathcal{X} = \{\mathrm{X}_n\}_{n=0}^{\mathrm{N}}$: a series of hidden states where each $\mathrm{X}_n \in \mathbb{R}^{\mathrm{D}}$ and X_0 is an (hypothetical) initial state.
- $\mathcal{Y} = \{\mathrm{Y}_n\}_{n=1}^{\mathrm{N}}$: a series of observations where each $\mathrm{Y}_n \in \mathbb{R}^{\mathrm{L}}$.
- $\mathcal{Z} = \{\mathrm{Z}_n\}_{n=1}^{\mathrm{N}}$: a series of latent states where $\mathrm{Z}_n \in \Theta$ could be either discrete or continuous.
- $\mathcal{F} = \{\mathrm{F}_\theta\}_{\theta\in\Theta}$: the parametrized transition model where F_θ is a $\mathrm{D} \times \mathrm{D}$ matrix.
- $\mathcal{H} = \{\mathrm{H}_\theta\}_{\theta\in\Theta}$: the parametrized emission model where H_θ is a $\mathrm{L} \times \mathrm{D}$ matrix.
- $\mathcal{Q} = \{\mathrm{Q}_\theta\}_{\theta\in\Theta}$: the parametrized covariance matrix in transition models where each Q_θ is a $\mathrm{D} \times \mathrm{D}$ matrix.
- $\mathcal{B} = \{b_\theta\}_{\theta\in\Theta}$: the parametrized bias vector in transition models where each $b_\theta \in \mathbb{R}^{\mathrm{D}}$.
- $\mathcal{R} = \{\mathrm{R}_\theta\}_{\theta\in\Theta}$: the parametrized covariance matrix in emission models where each R_θ is a $\mathrm{L} \times \mathrm{L}$ matrix.
- $\mathcal{C} = \{c_\theta\}_{\theta\in\Theta}$: the parametrized bias vector in emission models where each $c_\theta \in \mathbb{R}^{\mathrm{L}}$.

The major difference from the vanilla version in section A.2.1 and A.2.2 is that the models are now parametrized by latent states, as follows:

$$\begin{aligned} p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1},\mathrm{Z}_n\right) &= \mathcal{N}\left(\mathrm{X}_n|\mathrm{F}_{\mathrm{Z}_n}\mathrm{X}_{n-1}+b_{\mathrm{Z}_n},\mathrm{Q}_{\mathrm{Z}_n}\right) \\ p_e\left(\mathrm{Y}_n|\mathrm{X}_n,\mathrm{Z}_n\right) &= \mathcal{N}\left(\mathrm{Y}_n|\mathrm{H}_{\mathrm{Z}_n}\mathrm{X}_n+c_{\mathrm{Z}_n},\mathrm{R}_{\mathrm{Z}_n}\right) \end{aligned} \tag{A.34}$$

Note that each Z_n is a latent state thus not given a priori. The intuition comes from the following marginal transition and emission models that are expressed as mixtures of Gaussian:

$$\begin{aligned} p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right) &= \int \mathcal{N}\left(\mathrm{X}_n|\mathrm{F}_\theta\mathrm{X}_{n-1}+b_\theta,\mathrm{Q}_\theta\right)\pi\left(\theta\right)d\theta \\ p_e\left(\mathrm{Y}_n|\mathrm{X}_n\right) &= \int \mathcal{N}\left(\mathrm{Y}_n|\mathrm{H}_\theta\mathrm{X}_n+c_\theta,\mathrm{R}_\theta\right)\pi\left(\theta\right)d\theta \end{aligned} \tag{A.35}$$

, where π is a prior on θ. Thus, this model can be applied as an approximation method to non-Gaussian cases. The goal is to obtain the forward posterior $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ for each n. Suppose that $p\left(\mathrm{X}_{n-1}|\mathrm{Y}_{1:n-1}\right)$ is given as follows:

$$\begin{aligned} p\left(\mathrm{X}_{n-1}\middle|\mathrm{Y}_{1:n-1},\widetilde{\mathrm{Z}}_{1:n-1}^{(m)}\right) &= \mathcal{N}\left(\mathrm{X}_{n-1}\middle|\mu_{n-1}^{(m)},\Sigma_{n-1}^{(m)}\right) \\ p\left(\mathrm{X}_{n-1}|\mathrm{Y}_{1:n-1}\right) &= \sum_{m=1}^{\mathrm{M}}\omega_{n-1}^{(m)}\mathcal{N}\left(\mathrm{X}_{n-1}\middle|\mu_{n-1}^{(m)},\Sigma_{n-1}^{(m)}\right) \end{aligned} \tag{A.36}$$

, where $\widetilde{\mathrm{Z}}_{1:n-1}^{(m)}$ is the mth sample path of $\mathrm{Z}_{1:n-1}$.

Then, one can derive the following approximation to $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ by the importance sampling

with a proposal q:

$$
\begin{aligned}
p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) &= \iint \frac{p\left(\mathrm{X}_n, \mathrm{X}_{n-1}, \mathrm{Z}_n, \mathrm{Y}_n|\mathrm{Y}_{1:n-1}\right)}{p\left(\mathrm{Y}_n|\mathrm{Y}_{1:n-1}\right)} d\mathrm{Z}_n d\mathrm{X}_{n-1} \\
&= \sum_{m=1}^{\mathrm{M}} \omega_n^{(m)} \mathcal{N}\left(\mathrm{X}_n \middle| \mu_n^{(m)}, \Sigma_n^{(m)}\right) \\
&\propto \sum_{m=1}^{\mathrm{M}} \omega_{n-1}^{(m)} \int \mathcal{N}\left(\mathrm{X}_{n-1} \middle| \mu_{n-1}^{(m)}, \Sigma_{n-1}^{(m)}\right) \int p_e\left(\mathrm{Y}_n|\mathrm{X}_n, \mathrm{Z}_n\right) p_t\left(\mathrm{X}_n|\mathrm{X}_{n-1}, \mathrm{Z}_n\right) \pi\left(\mathrm{Z}_n\right) d\mathrm{Z}_n d\mathrm{X}_{n-1} \\
&\approx \sum_{m=1}^{\mathrm{M}} \omega_{n-1}^{(m)} \int \mathcal{N}\left(\mathrm{X}_{n-1} \middle| \mu_{n-1}^{(m)}, \Sigma_{n-1}^{(m)}\right) \frac{p_e\left(\mathrm{Y}_n \middle| \mathrm{X}_n, \widetilde{\mathrm{Z}}_n^{(m)}\right) p_t\left(\mathrm{X}_n \middle| \mathrm{X}_{n-1}, \widetilde{\mathrm{Z}}_n^{(m)}\right) \pi\left(\widetilde{\mathrm{Z}}_n^{(m)}\right)}{q\left(\widetilde{\mathrm{Z}}_n^{(m)} \middle| \widetilde{\mathrm{Z}}_{1:n-1}^{(m)}, \mu_{n-1}^{(m)}, \Sigma_{n-1}^{(m)}\right)} d\mathrm{X}_{n-1}
\end{aligned}
\tag{A.37}
$$

, where $\widetilde{\mathrm{Z}}_n^{(m)}$ is a sample drawn from the proposal $q\left(\mathrm{Z}_n \middle| \widetilde{\mathrm{Z}}_{1:n-1}^{(m)}, \mu_{n-1}^{(m)}, \Sigma_{n-1}^{(m)}\right)$ and:

$$
\begin{aligned}
\mu_{n|n-1}^{(m)} &\triangleq \mathrm{F}_{\widetilde{\mathrm{Z}}_n^{(m)}} \mu_{n-1}^{(m)} + b_{\widetilde{\mathrm{Z}}_n^{(m)}} \\
\Sigma_{n|n-1}^{(m)} &\triangleq Q_{\widetilde{\mathrm{Z}}_n^{(m)}} + \mathrm{F}_{\widetilde{\mathrm{Z}}_n^{(m)}} \Sigma_{n-1}^{(m)} \mathrm{F}_{\widetilde{\mathrm{Z}}_n^{(m)}}^{\mathbb{T}} \\
\mu_n^{(m)} &= \mu_{n|n-1}^{(m)} + \Sigma_{n|n-1}^{(m)} \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}}^{\mathbb{T}} \left(\mathrm{R}_{\widetilde{\mathrm{Z}}_n^{(m)}} + \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}} \Sigma_{n|n-1}^{(m)} \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}}^{\mathbb{T}}\right)^{-1} \left(\mathrm{Y}_n - \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}} \mu_{n|n-1}^{(m)} - c_{\widetilde{\mathrm{Z}}_n^{(m)}}\right) \\
\Sigma_n^{(m)} &= \Sigma_{n|n-1}^{(m)} - \Sigma_{n|n-1}^{(m)} \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}}^{\mathbb{T}} \left(\mathrm{R}_{\widetilde{\mathrm{Z}}_n^{(m)}} + \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}} \Sigma_{n|n-1}^{(m)} \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}}^{\mathbb{T}}\right)^{-1} \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}} \Sigma_{n|n-1}^{(m)} \\
\omega_n^{(m)} &\propto \frac{\omega_{n-1}^{(m)} \pi\left(\widetilde{\mathrm{Z}}_n^{(m)}\right) / q\left(\widetilde{\mathrm{Z}}_n^{(m)} \middle| \widetilde{\mathrm{Z}}_{1:n-1}^{(m)}, \mu_{n-1}^{(m)}, \Sigma_{n-1}^{(m)}\right)}{\mathcal{N}\left(\mathrm{Y}_n \middle| \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}} \mu_{n|n-1}^{(m)} + c_{\widetilde{\mathrm{Z}}_n^{(m)}}, \mathrm{R}_{\widetilde{\mathrm{Z}}_n^{(m)}} + \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}} \Sigma_{n|n-1}^{(m)} \mathrm{H}_{\widetilde{\mathrm{Z}}_n^{(m)}}^{\mathbb{T}}\right)}
\end{aligned}
\tag{A.38}
$$

Thus, one can approximate each $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ with a Gaussian mixture recursively by (A.37). Resampling on $\left\{\widetilde{\mathrm{Z}}_{1:n}^{(m)}\right\}_{m=1}^{\mathrm{M}}$ with respect to their weights $\left\{\omega_n^{(m)}\right\}_{m=1}^{\mathrm{M}}$ is recommended for stability. [Chen & Liu (2000)] also derive the extended version of the model which is not covered in this subsection.

There are other variants that deal with nonlinearity or non-Gaussian model aspects, as suggested in [Murphy (2012)]: extended Kalman filter (ENK) [McElhoe (1966)], unscented Kalman filter (UKF) [Julier & Uhlmann (1997)], assumed density filtering (ADF) [Maybeck (1979)] and ensemble Kalman filter [Evensen (1994); Houtekamer & Mitchell (1998)]. Note that those models do not guarantee the optimal inference as well.

A.3 Particle Filter and Smoother

Though section A.2.3 introduced an approximation variant of the Kalman filter, it is not enough to cover all cases: for example, $\mathcal{Y}$ is defined on a finite set while $\mathcal{X}$ still take continuous values.

Let us go back to the forward posterior updating formulation that is analogous to the forward algorithm of HMMs:

$$p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:n+1}\right) = \frac{p\left(\mathrm{Y}_{n+1}|\mathrm{X}_{n+1}\right)}{p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} \int p\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) d\mathrm{X}_n \tag{A.39}$$

The main obstacle that prevents the exact formulation is that the integration in (A.39) cannot be done analytically in general. The Monte Carlo methods approximate general integrations and the importance sampling [Gelman et al. (2013)] reduces the variance that could be very high in the naïve version. The core idea is to bring a proposal distribution q that is believed to get high density near to the mode of the posterior distribution $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}, \mathrm{X}_{n+1}\right)$ and approximate the integration as follows:

$$\int p\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) d\mathrm{X}_n \approx \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \frac{p\left(\mathrm{X}_{n+1}\middle|\widetilde{\mathrm{X}}_n^{(m)}\right) p\left(\widetilde{\mathrm{X}}_n^{(m)}\middle|\mathrm{Y}_{1:n}\right)}{q\left(\widetilde{\mathrm{X}}_n^{(m)}\right)} \tag{A.40}$$

, where $\widetilde{\mathrm{X}}_n^{(m)}$'s are independent and identically distributed samples drawn from q.

The particle filter (section A.3.1) and smoother (section A.3.2) [Doucet et al. (2001); Klaas et al. (2006)] are variational algorithms based on the sequential importance sampling. While the particle filter aims at updating the forward posterior $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ at each n, the particle smoother is for computing the posterior $p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right)$ after observing all $\mathrm{Y}_{1:\mathrm{N}}$.

A.3.1 Particle Filter

As section A.2.1, let us first define the following notations that will be used in section A.3.

- $\mathcal{X} = \{\mathrm{X}_n\}_{n=0}^{\mathrm{N}}$: a series of hidden states where each X_n and X_0 is an (hypothetical) initial

state.

- $\mathcal{Y} = \{\mathrm{Y}_n\}_{n=1}^{\mathrm{N}}$: a series of observations where each Y_n.

One major difference of the above definitions from those in section A.2.1 is that now X_n and Y_n are not restricted to be finite-dimensional vectors. Also, the transition and emission models $p_t^{(n)}\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right)$ and $p_e^{(n)}\left(\mathrm{Y}_n|\mathrm{X}_n\right)$ need not to be Gaussian here. In section A.2.1, each $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ is expressed as a Gaussian distribution that can be updated one-by-one recursively; section A.2.3 approximates it with a Gaussian mixture. Here, suppose that $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ is approximated with a weighted empirical density [Wasserman (2006)] as follows:

$$p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) = \sum_{m=1}^{\mathrm{M}} \omega_n^{(m)} 1_{\left\{\mathrm{X}_n = x_n^{(m)}\right\}}\left(\mathrm{X}_n\right) \tag{A.41}$$

The updating formulation depends on the proposal distribution q to sample X_{n+1}. We have the following:

$$\begin{aligned}
p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:n+1}\right) &= \int \frac{p_e^{(n+1)}\left(\mathrm{Y}_{n+1}|\mathrm{X}_{n+1}\right) p_t^{(n+1)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)}{p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} d\mathrm{X}_n \\
&= \frac{p_e^{(n+1)}\left(\mathrm{Y}_{n+1}|\mathrm{X}_{n+1}\right)}{p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} \int p_t^{(n+1)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) d\mathrm{X}_n \\
&= \frac{p_e^{(n+1)}\left(\mathrm{Y}_{n+1}|\mathrm{X}_{n+1}\right)}{p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} \sum_{k=1}^{\mathrm{M}} \int p_t^{(n+1)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) \omega_n^{(k)} 1_{\left\{\mathrm{X}_n = x_n^{(k)}\right\}}\left(\mathrm{X}_n\right) d\mathrm{X}_n \\
&= \frac{1}{p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} \sum_{k=1}^{\mathrm{M}} \omega_n^{(k)} p_e^{(n+1)}\left(\mathrm{Y}_{n+1}|\mathrm{X}_{n+1}\right) p_t^{(n+1)}\left(\mathrm{X}_{n+1}\middle|\mathrm{X}_n = x_n^{(k)}\right) \\
&= \sum_{k=1}^{\mathrm{M}} \frac{\omega_n^{(k)} p_e^{(n+1)}\left(\mathrm{Y}_{n+1}|\mathrm{X}_{n+1}\right) p_t^{(n+1)}\left(\mathrm{X}_{n+1}\middle|\mathrm{X}_n = x_n^{(k)}\right)}{q\left(\mathrm{X}_{n+1}\right) p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} q\left(\mathrm{X}_{n+1}\right)
\end{aligned} \tag{A.42}$$

The core idea is to approximate the proposal distribution q by the following empirical density estimation:

$$q\left(\mathrm{X}_{n+1}\right) \approx \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} 1_{\left\{\mathrm{X}_{n+1} = x_{n+1}^{(m)}\right\}}\left(\mathrm{X}_{n+1}\right) \tag{A.43}$$

, where $x_{n+1}^{(m)}$'s are the independent and identically distributed samples drawn from q. Here we

particularly call those samples "particles" that the algorithm is named after.

With (A.43), the below formulation follows:

$$
\begin{aligned}
p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:n+1}\right) &= \sum_{m=1}^{\mathrm{M}} \omega_{n+1}^{(m)} 1_{\left\{\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right\}}\left(\mathrm{X}_{n+1}\right) \\
&\approx \sum_{k=1}^{\mathrm{M}} \frac{\omega_n^{(k)} p_e^{(n+1)}\left(\mathrm{Y}_{n+1}|\mathrm{X}_{n+1}\right) p_t^{(n+1)}\left(\mathrm{X}_{n+1}\middle|\mathrm{X}_n=x_n^{(k)}\right)}{q\left(\mathrm{X}_{n+1}\right) p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} 1_{\left\{\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right\}}\left(\mathrm{X}_{n+1}\right) \\
&\propto \sum_{m=1}^{\mathrm{M}} \sum_{k=1}^{\mathrm{M}} \frac{\omega_n^{(k)} p_e^{(n+1)}\left(\mathrm{Y}_{n+1}\middle|\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right) p_t^{(n+1)}\left(\mathrm{X}_{n+1}=x_{n+1}^{(m)}\middle|\mathrm{X}_n=x_n^{(k)}\right)}{q\left(\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right) p\left(\mathrm{Y}_{n+1}|\mathrm{Y}_{1:n}\right)} 1_{\left\{\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right\}}\left(\mathrm{X}_{n+1}\right)
\end{aligned}
\tag{A.44}
$$

, where the weights $\omega_{n+1}^{(m)}$'s satisfy the following:

$$
\omega_{n+1}^{(m)} \propto \frac{p_e^{(n+1)}\left(\mathrm{Y}_{n+1}\middle|\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right)}{q\left(\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right)} \left(\sum_{k=1}^{\mathrm{M}} \omega_n^{(k)} p_t^{(n+1)}\left(\mathrm{X}_{n+1}=x_{n+1}^{(m)}\middle|\mathrm{X}_n=x_n^{(k)}\right)\right) \tag{A.45}
$$

So far, we have obtained an approximation (A.44) of $p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:n+1}\right)$ that has the same form of $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ in (A.41). For each n, the particle filter draws particles and computes their weights to recursively approximate the forward posteriors as above.

A.3.2 Particle Smoother

The particle smoother is analogous to the Kalman smoother in section A.2.2. The main goal is to approximate the posterior $p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right)$ with a weighted empirical density, as usual. Note that the particle filter returns $p\left(\mathrm{X}_{\mathrm{N}}|\mathrm{Y}_{1:\mathrm{N}}\right)$. Let us assume that $p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:\mathrm{N}}\right)$ has the following approximation:

$$
p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:\mathrm{N}}\right) \approx \sum_{m=1}^{\mathrm{M}} \hat{\omega}_{n+1}^{(m)} 1_{\left\{\mathrm{X}_{n+1}=x_{n+1}^{(m)}\right\}}\left(\mathrm{X}_n\right) \tag{A.46}
$$

Then, one can derive the following relationship:

$$\begin{aligned} p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right) &= \int p\left(\mathrm{X}_n, \mathrm{X}_{n+1}|\mathrm{Y}_{1:\mathrm{N}}\right) d\mathrm{X}_{n+1} \\ &= \int p\left(\mathrm{X}_n|\mathrm{X}_{n+1}, \mathrm{Y}_{1:n}\right) p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:\mathrm{N}}\right) d\mathrm{X}_{n+1} \\ &= \int \frac{p_t^{(n)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)}{p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:n}\right)} p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:\mathrm{N}}\right) d\mathrm{X}_{n+1} \end{aligned} \tag{A.47}$$

$p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:\mathrm{N}}\right)$ is from (A.46) and the particle filter returns an approximation of $p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right)$ for each n. For $p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:n}\right)$,

$$\begin{aligned} p\left(\mathrm{X}_{n+1}|\mathrm{Y}_{1:n}\right) &= \int p_t^{(n)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) d\mathrm{X}_n \\ &\approx \int p_t^{(n)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) \sum_{m=1}^{\mathrm{M}} \omega_n^{(m)} 1_{\left\{\mathrm{X}_n = x_n^{(m)}\right\}}\left(\mathrm{X}_n\right) d\mathrm{X}_n \\ &= \sum_{m=1}^{\mathrm{M}} \omega_n^{(m)} p_t^{(n)}\left(\mathrm{X}_{n+1}\middle|\mathrm{X}_n = x_n^{(m)}\right) \end{aligned} \tag{A.48}$$

Thus, we have the following approximation again for n:

$$p\left(\mathrm{X}_n|\mathrm{Y}_{1:\mathrm{N}}\right) \approx \sum_{m=1}^{\mathrm{M}} \hat{\omega}_n^{(m)} 1_{\left\{\mathrm{X}_n = x_n^{(m)}\right\}}\left(\mathrm{X}_n\right) \tag{A.49}$$

, where the weights $\hat{\omega}_n^{(m)}$'s are defined as follows:

$$\hat{\omega}_n^{(m)} \propto \omega_n^{(m)} \left(\sum_{k=1}^{\mathrm{M}} \hat{\omega}_{n+1}^{(k)} \frac{p_t^{(n)}\left(\mathrm{X}_{n+1} = x_{n+1}^{(k)}\middle|\mathrm{X}_n = x_n^{(m)}\right)}{\sum_{l=1}^{\mathrm{M}} \omega_n^{(l)} p_t^{(n)}\left(\mathrm{X}_{n+1} = x_{n+1}^{(k)}\middle|\mathrm{X}_n = x_n^{(l)}\right)} \right) \tag{A.50}$$

Note that particles that were sampled in the particle filter are again used in the approximations. Thus, weights $\hat{\omega}_n^{(m)}$'s could be of high variance in practice.

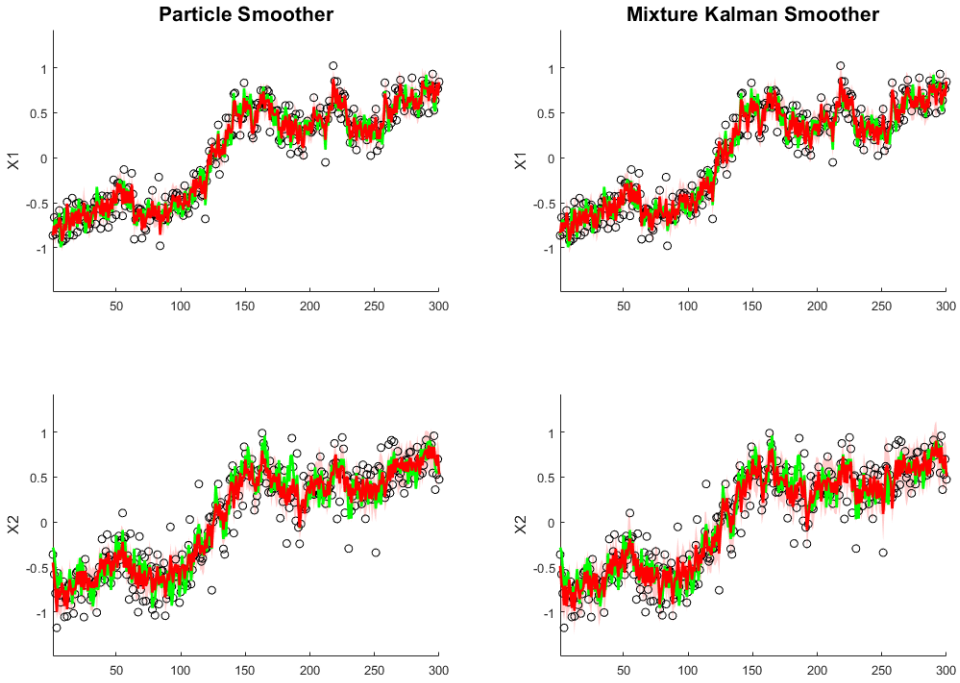


Figure A.6: Examples of the particle smoother and mixture Kalman smoother on the same 2-dimensional dataset. Black circles are the noisy observations and green and red curves are the true and inferred denoised hidden states, respectively.

Backward sampling directly comes from the following:

$$\begin{aligned} & p\left(\mathrm{X}_n|\mathrm{X}_{n+1}=\widetilde{x}_{n+1},\mathrm{Y}_{1:\mathrm{N}}\right)=p\left(\mathrm{X}_n|\mathrm{X}_{n+1}=\widetilde{x}_{n+1},\mathrm{Y}_{1:n}\right) \\ & \propto p_t^{(n)}\left(\mathrm{X}_{n+1}=\widetilde{x}_{n+1}|\mathrm{X}_n\right)p\left(\mathrm{X}_n|\mathrm{Y}_{1:n}\right) \\ & \approx \sum_{m=1}^{\mathrm{M}}\omega_n^{(m)}p_t^{(n)}\left(\mathrm{X}_{n+1}=\widetilde{x}_{n+1}\middle|\mathrm{X}_n=x_n^{(m)}\right)1_{\left\{\mathrm{X}_n=x_n^{(m)}\right\}}\left(\mathrm{X}_n\right) \end{aligned} \tag{A.51}$$

Thus, given $\mathrm{X}_{n+1}=\widetilde{x}_{n+1}$, X_n is drawn from the particles $\left\{x_n^{(m)}\right\}_{m=1}^{\mathrm{M}}$ with probabilities proportional to $\left\{\omega_n^{(m)}p_t^{(n)}\left(\mathrm{X}_{n+1}=\widetilde{x}_{n+1}\middle|\mathrm{X}_n=x_n^{(m)}\right)\right\}_{m=1}^{\mathrm{M}}$.

Figure A.6 shows examples of the particle smoother and mixture Kalman smoother applied to the same dataset. The performance of particle filtering (and smoother) is highly dependent on the choice of the proposal distribution q. If most of the particles bring 0 to the weights, the approximations on the forward posteriors are eventually controlled by only a small number of particles. Thus, the proposal distribution is often designed as a conditional distribution given

the particles at the previous indices and the associated observations. The resampling step [Murphy (2012)] supplements the algorithmic efficiency. However, all these attempts become to no purpose as the dimensionality of hidden states goes higher. Note that increasing the number of particles soon becomes intractable because the time complexity is quadratic to it, thus is not an ultimate solution. The Equivalent-weights particle filter [Ades & van Leeuwen (2013)] and local particle filter [Penny & Miyoshi (2016); Poterjoy (2016)] challenge such drawbacks.

A.4 Markov-chain Monte Carlo Algorithm

Section A.3 discussed about variational approximation methods for filtering and smoothing that can be applied in any transition and emission models. However, the sampling efficiency of particle smoother is still an unresolved challenge and, moreover, they do not return exact posteriors and sample paths but their approximation.

One might consider Markov-chain Monte Carlo (MCMC) [Liu (2001); Peters (2008)] algorithms that generate a series of sample paths from the given distributions (that have been either normalized or not). Under regular conditions, those sample paths converge to the exact ones, though the convergence is slow in practice especially if the initial sample path is in a region of low density. Because hidden states in state-space models are usually highly correlated one another, applying MCMC alone to the state-space models is not enough in general, but instead it can be utilized to supplement the particle filter and smoother for the refinement purpose: initializing sample paths by the particle smoother, and then running a MCMC algorithm from them. To avoid the dependency on the sample paths drawn from the MCMC algorithm, preparing multiple initial sample paths by the particle smoother is recommended.

There are various MCMC algorithms: this section briefly covers three of them, Gibbs sampling (section A.4.1), Metropolis-Hastings algorithm (section A.4.2) and Hamiltonian Monte Carlo algorithm (section A.4.3). The whole section A.4 take the notations in section A.3.

A.4.1 Gibbs Sampling

The basic idea of Gibbs sampling algorithm [Geman & Geman (1984)] is to sample each hidden state given all the others in turn. This algorithm is good if sampling each one given the others can be done efficiently while drawing together from the full posterior is not tractable. In mathematical terminology, X_n is drawn from the element-wise posterior $p\left(\mathrm{X}_n|\mathrm{X}_{\sim n}, \mathrm{Y}_{1:\mathrm{N}}\right)$ for each n. Note that one does not need to compute the full posterior one-by-one by considering the structure of state-space models as follows:

$$\begin{aligned} p\left(\mathrm{X}_n|\mathrm{X}_{\sim n}, \mathrm{Y}_{1:\mathrm{N}}\right) &\propto p\left(\mathrm{X}_n, \mathrm{Y}_n, \mathrm{X}_{n+1}|\mathrm{X}_{1:n-1}, \mathrm{X}_{n+2:\mathrm{N}}, \mathrm{Y}_{\sim n}\right) \\ &= p_e^{(n)}\left(\mathrm{Y}_n|\mathrm{X}_n\right) p_t^{(n)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n\right) p_t^{(n)}\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right) \end{aligned} \tag{A.52}$$

The Gibbs sampling algorithm starts with an initial path $\widetilde{\mathcal{X}} = \widetilde{\mathcal{X}}^{(0)} = \widetilde{\mathrm{X}}_{1:\mathrm{N}}^{(0)}$. Then it iterates the following for multiple times expected to cover all $\{1, 2, \cdots, \mathrm{N}\}$: firstly, sample n from $\{1, 2, \cdots, \mathrm{N}\}$ randomly, then sample X_n from $p\left(\mathrm{X}_n \middle| \widetilde{\mathcal{X}}_{\sim n}, \mathrm{Y}_{1:\mathrm{N}}\right)$ given $\widetilde{\mathcal{X}} = \widetilde{\mathcal{X}}^{(t)}$ and replace $\widetilde{\mathcal{X}}_n$ with it. After that, define the next $\widetilde{\mathcal{X}}^{(t+1)} = \widetilde{\mathcal{X}}$ and go back to the first step for obtaining a series $\left\{\widetilde{\mathcal{X}}^{(0)}, \widetilde{\mathcal{X}}^{(1)}, \cdots, \widetilde{\mathcal{X}}^{(\mathrm{T})}\right\}$ of T until the series has "burned in". Finally, return $\widetilde{\mathcal{X}}^{(\mathrm{T})}$ as a random sample from $p\left(\mathcal{X}|\mathcal{Y}\right)$ and go back to the very previous step with $\widetilde{\mathcal{X}} = \widetilde{\mathcal{X}}^{(\mathrm{T})}$ if another sample path is needed.

Note that $p\left(\mathrm{X}_n|\mathrm{X}_{\sim n}, \mathrm{Y}_{1:\mathrm{N}}\right)$ in (A.52) is depending only on the adjacent neighborhoods X_{n+1}, X_{n-1} and Y_n. Thus, one can run the blocked Gibbs sampling that draws $p\left(\mathrm{X}_{\mathcal{E}}|\mathrm{X}_{\sim\mathcal{E}}, \mathrm{Y}_{1:\mathrm{N}}\right)$ simultaneously where $\mathcal{E}$ is a set of either odd or even indices. Though it does not reduce the (theoretical) time complexity, some programming languages, such as MATLAB, are designed to run faster on this blocked version.

However, Gibbs sampling is usually inefficient because (A.52) is not represented explicitly for general transition and emission models. Besides rejection sampling, a practical variant is to fix a finite set of candidate assignments to the hidden state (by discretization) a priori and then to compute (A.52) to those candidates one-by-one for assigning weights used for the discrete sampling:

it is then efficient because storing a table for transition and emission models is now available.

A.4.2 Metropolis-Hastings Algorithm

As mentioned in the previous subsection, Gibbs sampling could be inefficient if exact sampling from the element-wise posteriors is not expressed explicitly. The Metropolis-Hastings algorithm [Metropolis et al. (1953); Hastings (1970); Martino et al. (2015)] would be more efficient in such a case. It additionally requires a proposal distribution q for drawing a candidate (proposal state) that could replace the previous one: defining it to be conditioned on the previously sampled hidden state makes the algorithm more efficient in practice.

The Metropolis-Hastings algorithm starts with an initial path $\widetilde{\mathcal{X}} = \widetilde{\mathcal{X}}^{(0)} = \widetilde{\mathrm{X}}_{1:\mathrm{N}}^{(0)}$. The iteration consists of the following steps: it first samples a candidate $\mathcal{X}^*$ from $q\left(\mathcal{X}\middle|\widetilde{\mathcal{X}}\right)$ given $\widetilde{\mathcal{X}} = \widetilde{\mathcal{X}}^{(t)}$. Then compute the following acceptance ratio α:

$$\alpha = \min\left\{1, \frac{p\left(\mathcal{X}^*|\mathcal{Y}\right)}{p\left(\widetilde{\mathcal{X}}\middle|\mathcal{Y}\right)} \cdot \frac{q\left(\widetilde{\mathcal{X}}\middle|\mathcal{X}^*\right)}{q\left(\mathcal{X}^*\middle|\widetilde{\mathcal{X}}\right)}\right\} \tag{A.53}$$

Lastly, replace $\widetilde{\mathcal{X}}$ with $\mathcal{X}^*$ in probability α ("accept" $\mathcal{X}^*$ in probability α) and define the next $\widetilde{\mathcal{X}}^{(t+1)} = \widetilde{\mathcal{X}}$.

After repeating the above steps for T times to get a series (chain) $\left\{\widetilde{\mathcal{X}}^{(0)}, \widetilde{\mathcal{X}}^{(1)}, \cdots, \widetilde{\mathcal{X}}^{(T)}\right\}$ until the series has "burned in", return $\widetilde{\mathcal{X}}^{(T)}$ as the sample drawn from $p\left(\mathcal{X}|\mathcal{Y}\right)$. Multiple samples are drawn by repeating the above steps.

Under regular conditions on the proposal distribution q, the Markov chain constructed by the Metropolis-Hastings algorithm is irreducible and aperiodic thus has (and converges to) the unique stationary distribution by the Fundamental Theorem of Markov chains [Norris (1997)]. Note that this theorem does not offer the mixing time that determines the number of iterations until the series (chain) has burned in. A practical way is to define q to change (at most) only a small number of hidden states per iteration.

Note that it is not required to compute the ratio of $p\left(\mathcal{X}^*|\mathcal{Y}\right)$ and $p\left(\widetilde{\mathcal{X}}\middle|\mathcal{Y}\right)$ by computing

them exactly: if q only replaces at most one state per iteration just like Gibbs sampling, then the ratio is given as follows: let X_n is the target to sample.

$$\frac{p\left(\mathcal{X}^*|\mathcal{Y}\right)}{p\left(\widetilde{\mathcal{X}}\middle|\mathcal{Y}\right)} = \frac{p_e^{(n)}\left(\mathrm{Y}_n|\mathrm{X}_n^*\right)p_t^{(n)}\left(\mathrm{X}_{n+1}|\mathrm{X}_n^*\right)p_t^{(n)}\left(\mathrm{X}_n^*|\mathrm{X}_{n-1}\right)}{p_e^{(n)}\left(\mathrm{Y}_n\middle|\widetilde{\mathrm{X}}_n\right)p_t^{(n)}\left(\mathrm{X}_{n+1}\middle|\widetilde{\mathrm{X}}_n\right)p_t^{(n)}\left(\widetilde{\mathrm{X}}_n\middle|\mathrm{X}_{n-1}\right)} \tag{A.54}$$

A similar simplification could also be obtained in computing the ratio of $q\left(\widetilde{\mathcal{X}}\middle|\mathcal{X}^*\right)$ and $q\left(\mathcal{X}^*\middle|\widetilde{\mathcal{X}}\right)$. If q is moreover designed to be symmetric, i.e., $q\left(\widetilde{\mathcal{X}}\middle|\mathcal{X}^*\right) \equiv q\left(\mathcal{X}^*\middle|\widetilde{\mathcal{X}}\right)$, then it is only required to compute (A.54) for obtaining the acceptance ratio α in (A.53). Note that Gibbs sampling is a special case of the Metropolis-Hastings algorithm where $q\left(\widetilde{\mathcal{X}}\middle|\mathcal{X}^*\right) = p\left(\widetilde{\mathcal{X}}\middle|\mathcal{Y}\right)$, which results in $\alpha \equiv 1$, i.e., always accept the candidate to replace the previous one.

A.4.3 Hamiltonian Monte Carlo Algorithm

One drawback of the Metropolis-Hastings algorithm is that the proposal distribution q affects the mixing time. Thus, it is natural to consider about the improvements involved with efficient sampling, or more comprehensively, better chains. The Hamiltonian (Hybrid) Monte Carlo (HMC) algorithm [Duane et al. (1987); Neal (1993); MacKay (2003); Neal (2010); Girolami & Calderhead (2011)] is such an outcome.

An HMC regards the parameters and hidden states as a "particle" and assumes auxiliary variables that represent the momentum of that particle. It updates not only the hidden states (as the models in sections A.4.1 and A.4.2) but also the momentum. Thus, the HMC requires the gradient of the log-posterior density of hidden states, better in the computational efficiency if it is analytically computable. The following steps are referring to [Gelman et al. (2013)].

Step 0. Initialization:

Step 0.1. Fix a mass covariance matrix Σ and learning rate $\varepsilon > 0$.

Step 0.2. Draw the momentum vector $\mathcal{Z}^{(0)}$ from $\mathcal{N}\left(\vec{0}, \Sigma\right)$ and the hidden state $\mathcal{X}^{(0)}$.

Step 1. Repeat the following for T times, until the burn-in status is reached:

Step 1.1. Initialize $\left(\mathcal{Z}^{(t,0)}, \mathcal{X}^{(t,0)}\right) = \left(\mathcal{Z}^{(t)}, \mathcal{X}^{(t)}\right)$ and repeat the following 'leapfrog' steps for

L times:

Step 1.1.1. Update the momentum vector by following:

$$\mathcal{Z}^{(t,l+0.5)} = \mathcal{Z}^{(t,l)} + \frac{1}{2}\varepsilon \cdot \nabla_{\mathcal{X}} \log p\left(\mathcal{X}^{(t,l)}|\mathcal{Y}\right) \tag{A.55}$$

Step 1.1.2. Update the hidden states by following:

$$\mathcal{X}^{(t,l+1)} = \mathcal{X}^{(t,l)} + \varepsilon \cdot \Sigma^{-1}\mathcal{Z}^{(t,l+0.5)} \tag{A.56}$$

Step 1.1.3. Update the momentum vector again by following:

$$\mathcal{Z}^{(t,l+1)} = \mathcal{Z}^{(t,l+0.5)} + \frac{1}{2}\varepsilon \cdot \nabla_{\mathcal{X}} \log p\left(\mathcal{X}^{(t,l+1)}|\mathcal{Y}\right) \tag{A.57}$$

Step 1.2. Compute the following acceptance rate r:

$$r = \min\left\{1, \frac{p\left(\mathcal{X}^{(t,\mathrm{L})}|\mathcal{Y}\right)\mathcal{N}\left(\mathcal{Z}^{(t,\mathrm{L})}|\vec{0},\Sigma\right)}{p\left(\mathcal{X}^{(t)}|\mathcal{Y}\right)\mathcal{N}\left(\mathcal{Z}^{(t)}|\vec{0},\Sigma\right)}\right\} \tag{A.58}$$

Step 1.3. Define $\left(\mathcal{X}^{(t+1)}, \mathcal{Z}^{(t+1)}\right) = \left(\mathcal{X}^{(t,\mathrm{L})}, \mathcal{Z}^{(t,\mathrm{L})}\right)$ with probability r.

Otherwise, $\left(\mathcal{X}^{(t+1)}, \mathcal{Z}^{(t+1)}\right) = \left(\mathcal{X}^{(t)}, \mathcal{Z}^{(t)}\right)$.

Step 2. Return $\mathcal{X}^{(\mathrm{T})}$ as the sample path drawn from $p\left(\mathcal{X}|\mathcal{Y}\right)$.

Step 3. Go back to step 1 until enough sample paths are drawn.

Because $\log p\left(\mathcal{X}|\mathcal{Y}\right) = \log p\left(\mathcal{Y}|\mathcal{X}\right) + \log p\left(\mathcal{X}\right) - \log p\left(\mathcal{Y}\right)$ and $\nabla_{\mathcal{X}} \log p\left(\mathcal{Y}\right) \equiv 0$, one can compute $\nabla_{\mathcal{X}} \log p\left(\mathcal{X}|\mathcal{Y}\right)$ as follows:

$$\begin{aligned}\nabla_{\mathcal{X}} \log p\left(\mathcal{X}|\mathcal{Y}\right) &= \nabla_{\mathcal{X}} \log p\left(\mathcal{Y}|\mathcal{X}\right) + \nabla_{\mathcal{X}} \log p\left(\mathcal{X}\right) \\ &= \sum_{n=1}^{\mathrm{N}} \nabla_{\mathcal{X}} \log p_e^{(n)}\left(\mathrm{Y}_n|\mathrm{X}_n\right) + \sum_{n=1}^{\mathrm{N}} \nabla_{\mathcal{X}} \log p_t^{(n)}\left(\mathrm{X}_n|\mathrm{X}_{n-1}\right)\end{aligned} \tag{A.59}$$

Thus, it is not required to compute $\nabla_{\mathcal{X}} \log p\left(\mathcal{X}|\mathcal{Y}\right)$ from $p\left(\mathcal{X}|\mathcal{Y}\right)$ directly but just enough

to compute the two sums in the right-hand of (A.59).

If both transition and emission models are analytically differentiable, computing (A.59) can be done efficiently. Also, giving a diagonal mass covariance matrix allows to compute (A.56) fast. [Hoffman & Gelman (2014)] suggests a way of choosing the number of steps L and learning rate ε which the performance is sensitive to automatically.

However, the HMC is a bit limited in practice. It cannot be applied to the state-space model when $\mathcal{X}$ is defined in a discrete space. Strictly positive posteriors are required because it depends on the gradient of log-posteriors. Clearly, it is prohibited if either the transition or emission model is nondifferentiable. For example, the core alignment algorithm in section 2.2 has a non-differentiable transition model thus the HMC cannot be applied to that problem. Despite of such limitation, the HMC is a good variant of the Metropolis-Hastings for dealing with high-dimensional and highly correlated hidden states.

B

Gaussian Process Models

The Gaussian Process (GP) [Rasmussen & Williams (2006)] is a stochastic process where any finite collection of random variables follows a (multivariate) Gaussian distribution. More formally, let $\mathcal{X} = \{\mathrm{X}_t\}_{t\in\mathcal{T}}$ be a stochastic process defined over a set of indices $\mathcal{T}$. Then, $\mathcal{X}$ is a GP if and only if $\mathrm{X_T} \triangleq (\mathrm{X}_{t_n})_{n=1}^{\mathrm{N}}$ is a multivariate Gaussian random variable for any finite subset $\mathrm{T} = \{t_n\}_{n=1}^{\mathrm{N}} \subseteq \mathcal{T}$. More constructively, a GP is entirely specified by its mean function μ and covariance function $\mathbb{K}$ that are defined on $\mathcal{T}$ and $\mathcal{T}\times\mathcal{T}$, respectively: $p(\mathrm{X_T}) = \mathcal{N}\left(\mathrm{X_T}\middle|\mu_{\mathrm{T}}, \mathbb{K}_{\mathrm{TT}}\right)$: it is sometimes denoted by $\mathcal{GP}\left(\mathrm{X_T}\middle|\mu_{\mathrm{T}}, \mathbb{K}_{\mathrm{TT}}\right)$, to emphasize that $\mathrm{X_T}$ follows a GP. Rest of the notations are defined in section 1.4.1. The covariance function $\mathbb{K}$ should be not only positive semi-definite and symmetric but also consistently extendable as T varies, thus a set of functions

that satisfy some certain conditions are considered: those are called 'kernel' covariance functions. Details can be found in section B.1.

In statistical learning, GPs are popular for nonparametric priors on the continuous random functions. Suppose that we have a set of inputs $\mathcal{X}$ and the associated outputs $\mathcal{Y}$ and want to give a relationship between $\mathcal{X}$ and $\mathcal{Y}$. Instead of constructing a direct function that maps $\mathcal{X}$ to $\mathcal{Y}$, $p(\mathcal{Y}|\mathcal{X})$, here we assume a latent state β that might explain $\mathcal{Y}$ more directly and then consider the relationship between $\mathcal{X}$ and β, $p(\beta|\mathcal{X})$, and $(\mathcal{X},\beta)$ and $\mathcal{Y}$, $p(\mathcal{Y}|\mathcal{X},\beta)$. Specifically, β follows a GP with $\mathcal{X}$ as an input: $p(\beta|\mathcal{X}) = \mathcal{GP}\left(\beta\middle|\mu_{\mathcal{X}}, \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)$. The goal is constructing a function which maps an arbitrary query input x to a distribution of the associated output y given $\mathcal{X}$ and $\mathcal{Y}$ by marginalizing the latent variables out, i.e.,

$$p(y|x,\mathcal{X},\mathcal{Y}) = \int p(y|x,b) \int p(b|\beta,x,\mathcal{X})\, p(\beta|\mathcal{X},\mathcal{Y})\, d\beta db \tag{B.1}$$

By assumption, (β, b) also follows a GP with $(\mathcal{X}, x)$ as an input. Thus, one can easily derive the following:

$$p(b|\beta,x,\mathcal{X}) = \mathcal{N}\left(b\middle|\mu_x + \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\left(\beta - \mu_{\mathcal{X}}\right), \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}x}\right) \tag{B.2}$$

$p(y|x,b)$ is defined by the assumed relationship $p(\mathcal{Y}|\mathcal{X},\beta)$, and $p(\beta|\mathcal{X},\mathcal{Y})$ is a posterior with $p(\beta|\mathcal{X})$ as the prior and $p(\mathcal{Y}|\mathcal{X},\beta)$ as the likelihood. The characteristics of a GP model are determined by the choice of $p(\mathcal{Y}|\mathcal{X},\beta)$, i.e., how outputs are explained by not only the associated inputs but also the latent state, and the choice of mean and kernel covariance function that define $p(\beta|\mathcal{X})$, which makes the model very flexible. Sections B.2 to B.5 cover the specific applications of the GP models.

Because the latent state is marginalized out, the final model is depending only on the training data and kernel covariance function, thus overfitting is not caused by the excessive number of parameters. The nonparametric aspect of GPs gives a structure-free inference. Tuning hyperparameters is often a convoluted issue in nonparametric modelling because the objective function

is not given certainly. However, the marginal likelihood (ML) of $\mathcal{Y}$ given $\mathcal{X}$, $p(\mathcal{Y}|\mathcal{X})$, is often chosen as the objective function to maximize in GP models for tuning kernel covariance functions.

There are two principal drawbacks of the GP models. One is that the model is often suboptimal due to the inflexible kernel covariance function [Remes et al. (2017); Blomqvist et al. (2019)]. The other is about the computational issue: the inference is involved with at least one matrix inversion $\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}$ of size $|\mathcal{X}| \times |\mathcal{X}|$ thus quickly becomes intractable ($\mathcal{O}\left(|\mathcal{X}|^{2.373}\right)$ [Le Gall (2014)]) as the size of training data $|\mathcal{X}|$ increases. Section B.6 introduces a deep Gaussian process to deal with the first issue and section B.7 shows a variational approximation method to address the second one.

B.1 Kernels

Kernel covariance functions (kernels) play a key role in the inference based on GPs: roughly speaking, kernels give some prior assumptions on the characteristics such as smoothness, non-stationary, periodicity, and multiscale or hierarchical structures [Gelman et al. (2013)]. Not every function can be a kernel. A formal definition is given as follows: a symmetric function $\mathbb{K} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is called a (positive semi-definite) kernel on a nonempty set $\mathcal{T}$ if the following holds for any $N \in \mathbb{N}$, $\{t_n\}_{n=1}^{\mathrm{N}} \subseteq \mathcal{T}$ and $\{c_n\}_{n=1}^{\mathrm{N}} \subseteq \mathbb{R}$:

$$\sum_{m=1}^{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} c_m c_n \mathbb{K}(t_m, t_n) \geq 0 \tag{B.3}$$

More intuitively, we can say that the kernel function $\mathbb{K}(u, v)$ measures how much the associated outputs X_u and X_v are correlated. Each kernel is parametrized by a small set of kernel hyperparameters thus 'tuning' kernels means to tune involved kernel hyperparameters. The misspecification problem stems from the facts that there are infinitely many kernels and no perfect and objective rule exists for the real data that potentially include outliers, though there have been some publications to deal with it: see [Duvenaud et al. (2013); Ben Abdessalem et al. (2019)] for

examples. Note that any sums and products of kernels are again kernels: sums are straightforward whereas products are by the Mercer's theorem.

Kernels can be defined on any set $\mathcal{T}$ (e.g. string kernels [Lodhi et al. (2002)] and fisher kernels [Jaakkola et al. (2000)]), but here we assume that the kernel functions are computed at vector realizations, i.e., $\mathcal{T} = \mathbb{R}^{\mathrm{D}}$ for some $\mathrm{D} \in \mathbb{N}$. The following subsections cover some popular kernels [Rasmussen & Williams (2006); Barber (2012); Duvenaud (2014)].

B.1.1 Stationary Kernels

A stationary kernel is a kernel function $\mathbb{K} : \mathbb{R}^{\mathrm{D}} \times \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ that satisfies $\mathbb{K}(u, v) \equiv f(u - v)$ for some function $f : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$. Thus, the functional value depends only on the relative difference between the inputs: it implies that the prior assumptions on the characteristics are universal in the domain.

It is natural to start with the most trivial one. A constant kernel $\mathbb{K}_{\mathrm{const}}$ with a kernel hyperparameter σ_0^2 is defined by following:

$$\mathbb{K}_{\mathrm{const}}(u, v) = \sigma_0^2 \cdot 1_{\{u=v\}} \tag{B.4}$$

, which implies that $\mathbb{K}_{\mathcal{X}\mathcal{X}} \equiv \sigma_0^2 \cdot \mathbb{I}$ and $\mathbb{K}_{x\mathcal{X}} \equiv 0$, so the $p(b|\beta, x, \mathcal{X})$ in (B.2) is reduced to the following:

$$p(b|\beta, x, \mathcal{X}) = \mathcal{N}\left(b\middle|\mu_x, \sigma_0^2\right) \tag{B.5}$$

Here (B.5) is independent of β and $\mathcal{X}$, so $p(y|x, \mathcal{X}, \mathcal{Y}) = \int p(y|x, b)\, \mathcal{N}\left(b\middle|\mu_x, \sigma_0^2\right) db$: a naïve Bayes modelling. Because any sum of kernels is again a kernel, constant kernels are usually added for the numerical stability and overlapping inputs.

Now let us discuss about less trivial cases. A squared-exponential (SE) kernel $\mathbb{K}_{\mathrm{SE}}$ with kernel hyperparameters σ_0^2 and $l > 0$ has the following form:

$$\mathbb{K}_{\mathrm{SE}}(u, v) = \sigma_0^2 \cdot \exp\left(-\frac{|u - v|^2}{2l^2}\right) \tag{B.6}$$

Here, σ_0^2 determines the average distance of the GP from its mean and the characteristic length-scale l scales the distance $|u - v|$. If $\mathrm{D} > 1$, adopting (B.6) is usually suboptimal: the natural extension is the following:

$$\mathbb{K}_{\mathrm{SE}}(u, v) = \sigma_0^2 \cdot \exp\left(-\frac{1}{2}\left|\mathrm{L}(u - v)\right|^2\right) \tag{B.7}$$

, where L is a $\mathrm{D} \times \mathrm{D}$ matrix.

The SE kernel makes the associated GP infinitely mean square (MS) differentiable thus the associated GP becomes very smooth, which makes it the most popular kernel in the field, though [Stein (1999)] argues that SE kernels are too smooth to model real physical processes. Note that the SE kernel is also called as the radial basis function (RBF) kernel.

In fact, the above SE kernel is a special case of the Matérn class of kernel functions [Matérn (1986); Stein (1999); Genton (2002)]. A Matérn kernel with kernel hyperparameters $\sigma_0^2, d > 0$ and $l > 0$ has the following form:

$$\mathbb{K}_{\mathrm{Matern}}^{(d)}(u, v) = \sigma_0^2 \cdot \frac{2^{1-d}}{\Gamma(d)}\left(\frac{\sqrt{2d}}{l}|u - v|\right)^d K_d\left(\frac{\sqrt{2d}}{l}|u - v|\right) \tag{B.8}$$

, where K_d is a modified Bessel function [Abramowitz & Stegun (1964)]. Each kernel is specified by d, and it can be shown that $\mathbb{K}_{\mathrm{Matern}}^{(d)}(u, v)$ converges to $\mathbb{K}_{\mathrm{SE}}(u, v)$ as $d \to \infty$ [Rasmussen & Williams (2006)].

The form in (B.8) can be simplified if $d = n - 1/2$ for $n \in \mathbb{N}$. The examples are following:

1. $d = 1/2$: (Ornstein-Uhlenbeck (OU))

$$\mathbb{K}_{\mathrm{OU}}(u, v) \triangleq \mathbb{K}_{\mathrm{Matern}}^{(1/2)}(u, v) = \sigma_0^2 \cdot \exp\left(-\frac{|u - v|}{l}\right) \tag{B.9}$$

2. $d = 3/2$:

$$\mathbb{K}_{\mathrm{Matern}}^{(3/2)}(u, v) = \sigma_0^2 \cdot \left(1 + \frac{\sqrt{3}|u - v|}{l}\right)\exp\left(-\frac{\sqrt{3}|u - v|}{l}\right) \tag{B.10}$$

3. $d = 5/2$:

$$\mathbb{K}_{\text{Matern}}^{(5/2)}(u, v) = \sigma_0^2 \cdot \left(1 + \frac{\sqrt{5}\,|u - v|}{l} + \frac{5\,|u - v|^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}\,|u - v|}{l}\right) \tag{B.11}$$

In practice, $\mathbb{K}_{\text{Matern}}^{(5/2)}(u, v)$ is quite similar with the SE kernel, thus $\mathbb{K}_{\text{Matern}}^{(d)}(u, v)$ for $d > 5/2$ is often not considered. A GP with the Matérn kernel $\mathbb{K}_{\text{Matern}}^{(d)}$ is k-times MS differentiable if and only if $k < d$.

As discussed in the SE kernel, a natural extension of the Matérn kernels for $\mathrm{D} > 1$ is given as follows for a $\mathrm{D} \times \mathrm{D}$ matrix L:

$$\mathbb{K}_{\text{Matern}}^{(d)}(u, v) = \sigma_0^2 \cdot \frac{2^{1-d}}{\Gamma(d)} \left(\sqrt{2d} \cdot |\mathrm{L}(u - v)|\right)^d K_d \left(\sqrt{2d} \cdot |\mathrm{L}(u - v)|\right) \tag{B.12}$$

One noteworthy characteristic of the Matérn kernels is that each of them is in connection with a continuous version of the autoregressive (AR) model [Stein (1999); Papoulis & Pillai (2001); Rasmussen & Williams (2006)] where $\mathbb{K}_{\text{Matern}}^{(d)}$ corresponds to AR $(d + 1/2)$. We will revisit this characteristic in section B.4.

The SE kernel is also in another family of kernels, the γ-exponential (GE) kernels [Schoenberg (1938); Stein (1999)], that with kernel hyperparameters σ_0^2, $0 < \gamma \leq 2$ and $l > 0$ have the following form:

$$\mathbb{K}_{\text{GE}}^{(\gamma)}(u, v) = \sigma_0^2 \cdot \exp\left(-\left(\frac{|u - v|}{\sqrt{2}l}\right)^\gamma\right) \tag{B.13}$$

Note that $\mathbb{K}_{\text{GE}}^{(2)}(u, v) \equiv \mathbb{K}_{\text{SE}}(u, v)$. One limitation of these kernels is that the corresponding GPs are not MS differentiable unless $\gamma = 2$. A natural extension for $\mathrm{D} > 1$ is given as follows for a $\mathrm{D} \times \mathrm{D}$ matrix L:

$$\mathbb{K}_{\text{GE}}^{(\gamma)}(u, v) = \sigma_0^2 \cdot \exp\left(-\left(\frac{|\mathrm{L}(u - v)|}{\sqrt{2}}\right)^\gamma\right) \tag{B.14}$$

The rational quadratic (RQ) kernel [Rasmussen & Williams (2006)] is derived by a scale mixture of SE kernels with a certain gamma prior on the inverse of the square of characteristic length-

scales: now $\sigma_0^2, \alpha > 0$ and $l > 0$ are the kernel hyperparameters.

$$\begin{aligned}
\mathbb{K}_{\mathrm{RQ}}^{(\alpha)}(u, v) &= \sigma_0^2 \int \Gamma\left(\tau | \alpha, \alpha l^2\right) \exp\left(-\frac{|u-v|^2}{2}\tau\right) d\tau \\
&= \sigma_0^2 \int \frac{\left(\alpha l^2\right)^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp\left(-\alpha l^2 \tau - \frac{|u-v|^2}{2}\tau\right) d\tau \\
&= \sigma_0^2 \left(1 + \frac{|u-v|^2}{2\alpha l^2}\right)^{-\alpha}
\end{aligned} \tag{B.15}$$

Note that $\mathbb{K}_{\mathrm{RQ}}^{(\alpha)}(u, v)$ converges to $\mathbb{K}_{\mathrm{SE}}(u, v)$ as $\alpha \to \infty$. One advantage of the RQ kernel over the Matérn and GE kernels shown above is that the corresponding GP is now infinitely MS differentiable for every α. One can show that the following variant for a D $\times$ D matrix L is also a valid kernel:

$$\mathbb{K}_{\mathrm{RQ}}^{(\alpha)}(u, v) = \sigma_0^2 \left(1 + \frac{|\mathrm{L}(u-v)|^2}{2\alpha}\right)^{-\alpha} \tag{B.16}$$

The kernels that have been covered so far are just stationary. [MacKay (1998)] derives the following periodic kernel which allows to model functions repeating themselves exactly:

$$\mathbb{K}_{\mathrm{Period}}^{(p)}(u, v) = \sigma_0^2 \cdot \exp\left(-\frac{2\sin^2(\pi |u-v|/p)}{l^2}\right) \tag{B.17}$$

Like the SE kernel, still σ_0^2 determines the average distance of the GP from its mean and the characteristic length-scale l scales the distance $|u-v|$. The period p decides the frequency of function. Note that the above periodic kernel is defined on $\mathbb{R} \times \mathbb{R}$ only. The extension to the multidimensional case is straightforwardly given in [Haji Ghassemi (2013)]:

$$\mathbb{K}_{\mathrm{Period}}^{(p)}(u, v) = \sigma_0^2 \cdot \exp\left(-\sum_{d=1}^{\mathrm{D}} \frac{2\sin^2(\pi |u_d - v_d|/p)}{l_d^2}\right) \tag{B.18}$$

Figure B.1 shows three functions that are drawn from the GP with each kernel described in this subsection. CONST, SE, OU, M15, M25, GE, RQ and PERIOD stand for the constant kernel, squared-exponential kernel, Ornstein-Uhlenbeck kernel, Matérn kernel with $d = 3/2$,

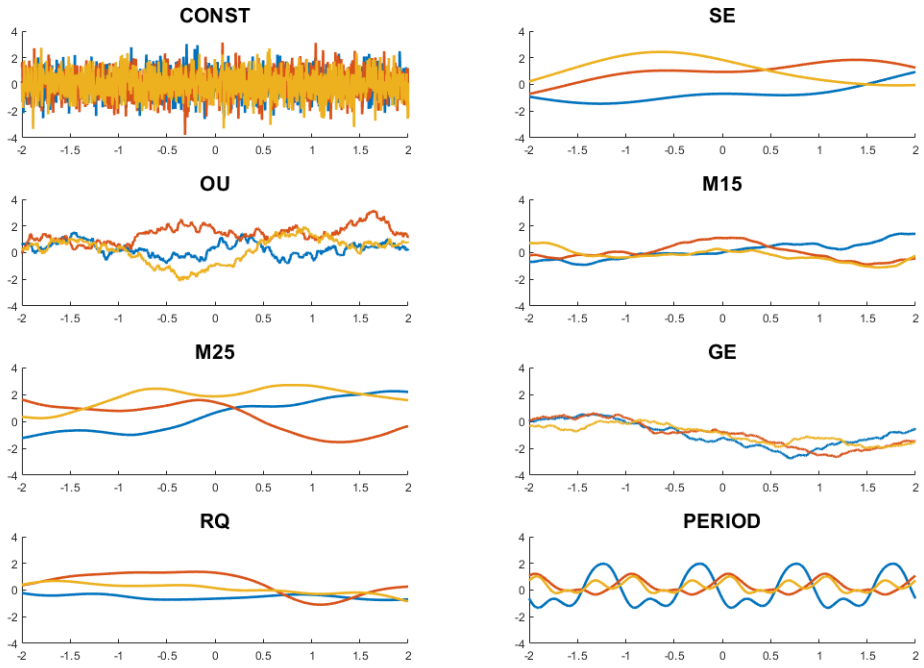


Figure B.1: Functions sampled from the Gaussian process with different stationary kernels.

Matérn kernel with $d = 5/2$, γ-exponential kernel with $\gamma = 1.5$, rational quadratic kernel with $\alpha = 0.5$ and periodic kernel with $p = 1$, respectively: we set $\sigma_0^2 = 1$ and $l = 1$ for each case.

B.1.2 Nonstationary Kernels

For a given kernel $\mathbb{K}(u, v)$ and a feature function $\varphi : \mathbb{R}^D \to \mathbb{R}^E$, $\widetilde{\mathbb{K}}(u, v) \triangleq \mathbb{K}(\varphi(u), \varphi(v))$ is again a kernel function, by definition. This feature function gives the GP models more flexibility. Nonstationary kernels are often adorned with feature functions because they are not needed to be dependent only on the input difference as the stationary cases. In this section, we assume that φ is an arbitrary feature function.

As we did in section B.1.1, let us start again with a trivial case. A linear kernel with Σ is defined by following:

$$\mathbb{K}_{\text{linear}}(u, v) = \varphi(u)^{\mathrm{T}} \Sigma \varphi(v) \tag{B.19}$$

, where Σ is a covariance function. Note that the GP regression with $\mathbb{K}_{\text{linear}}(u, v)$ results in a

linear regression with the generalized least squares (GLS) method.

Scaling on the stationary kernels on high-dimensional spaces is done on their input distances. A similar thing can be done in the nonstationary cases. A Gibbs kernel [Gibbs (1997)] is defined by following:

$$\mathbb{K}_{\text{Gibbs}}(u,v) = \prod_{d=1}^{\mathrm{D}} \left(\frac{2 l_d(u)\, l_d(v)}{l_d^2(u) + l_d^2(v)} \right)^{1/2} \cdot \exp\left(- \sum_{d=1}^{\mathrm{D}} \frac{(u_d - v_d)^2}{l_d^2(u) + l_d^2(v)} \right) \tag{B.20}$$

, where each l_d is a positive function on $\mathbb{R}^{\mathrm{D}}$ that scales inputs.

Lastly, neural network kernels are derived as limiting cases of (shallow) neural networks with infinite hidden units with particular activation functions: σ_0^2 and Σ are the kernel hyperparameters.

1. The erf activation function [Williams (1998)]:

$$\mathbb{K}_{\text{NN}}^{\text{erf}}(u,v) = \sigma_0^2 \cdot \sin^{-1}\left(\frac{2\varphi(u)^{\mathbb{T}} \Sigma \varphi(v)}{\sqrt{\left(1 + 2\varphi(u)^{\mathbb{T}} \Sigma \varphi(u)\right)\left(1 + 2\varphi(v)^{\mathbb{T}} \Sigma \varphi(v)\right)}} \right) \tag{B.21}$$

2. The step activation function [Cho & Saul (2009)]:

$$\mathbb{K}_{\text{NN}}^{\text{step}}(u,v) = \sigma_0^2 (\pi - \xi(u,v)) \tag{B.22}$$

3. The rectified linear unit (ReLU) activation function [Cho & Saul (2009)]:

$$\mathbb{K}_{\text{NN}}^{\text{ReLU}}(u,v) = \sigma_0^2 (\sin \xi(u,v) + (\pi - \xi) \cos \xi(u,v)) \tag{B.23}$$

, where:

$$\xi(u,v) \triangleq \cos^{-1}\left(\frac{\varphi(u)^{\mathbb{T}} \Sigma \varphi(v)}{\sqrt{\varphi(u)^{\mathbb{T}} \Sigma \varphi(u) \cdot \varphi(v)^{\mathbb{T}} \Sigma \varphi(v)}} \right) \tag{B.24}$$

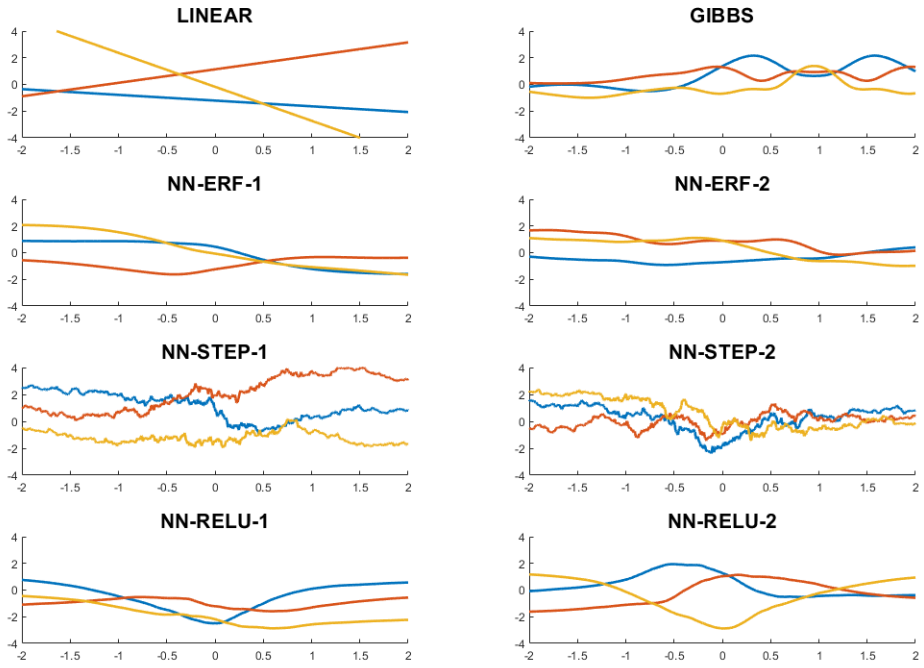


Figure B.2: Functions sampled from the Gaussian process with different nonstationary kernels.

In other words, some shallow neural networks that are involved with a lot of parameters can be converted to the GP models that are controlled by only a few hyperparameters.

Figure B.2 shows three functions that are drawn from the GP with each kernel described in this subsection. LINEAR, GIBBS, NN-ERF, NN-STEP, NN-RELU stand for the linear, Gibbs, neural network kernel with the erf activation function, neural network kernel with the step activation function and neural network kernel with ReLU activation function, respectively: we set $\sigma_0^2 = 1$ for each case. Two panels are assigned to each neural network kernel: the first one assumes $\varphi(u) = (1, u)$ and the second one takes $\varphi(u) = (1, u, u^2)$.

In applications, kernels are often designed by the combinations of the sums and products of the above stationary and nonstationary ones: constructing the "best" kernel for the training data is still an open problem. Choosing a proper feature function is another knotty issue: see [Calandra et al. (2016)].

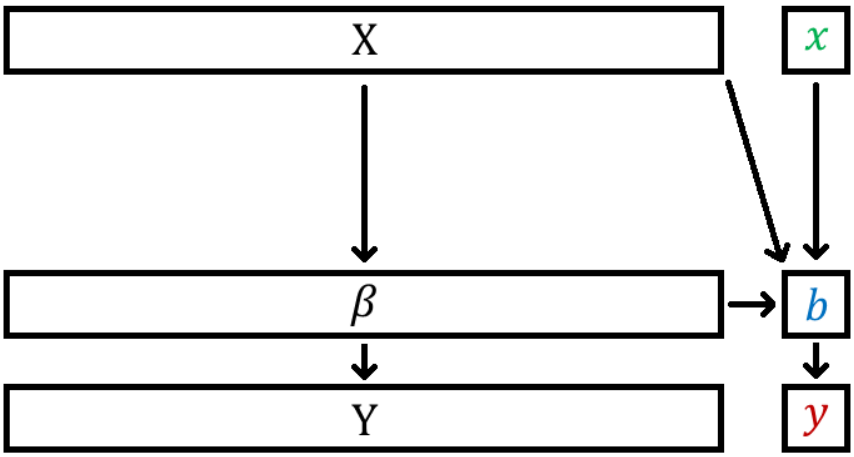


Figure B.3: A graphical representation of GPR and GPC. Capital X and Y are the training inputs and outputs, respectively, and β indicates the latent state, that is either regression (GPR) or log-odd (GPC). Green x and red y are the query input and the associated output, respectively, and blue b stands for the associated latent variable.

B.2 Gaussian Process Regression

Regression aims at finding a simpler model that describes the given data for the explanatory and/or predictive purposes. In particular, statistical regression constructs probabilistic models as the regression functions. Gaussian process regression (GPR) [Rasmussen & Williams (2006)] is a nonparametric regression method that reduces the training data to a probabilistic model: its graphical representation is given in figure B.3.

Let us recall the general setting at the beginning of this chapter: the GP model is determined by $p(\mathcal{Y}|\mathcal{X},\beta)$. A GPR assumes that the latent state β implies the regression of $\mathcal{Y}$, thus here $p(\mathcal{Y}|\mathcal{X},\beta)$ is called an observational model. For now, the observational model is assumed to be Gaussian and each $\mathrm{Y}_n \in \mathbb{R}$:

$$\begin{aligned} p(\mathcal{Y}|\mathcal{X},\beta) &= \mathcal{N}(\mathcal{Y}|\beta,\Lambda_{\mathcal{X}}) = \prod_{n=1}^{\mathrm{N}} \mathcal{N}\left(\mathrm{Y}_n|\beta_n,\Lambda_{\mathrm{X}_n}\right) \\ p(y|x,b) &= \mathcal{N}(y|b,\Lambda_x) \end{aligned} \tag{B.25}$$

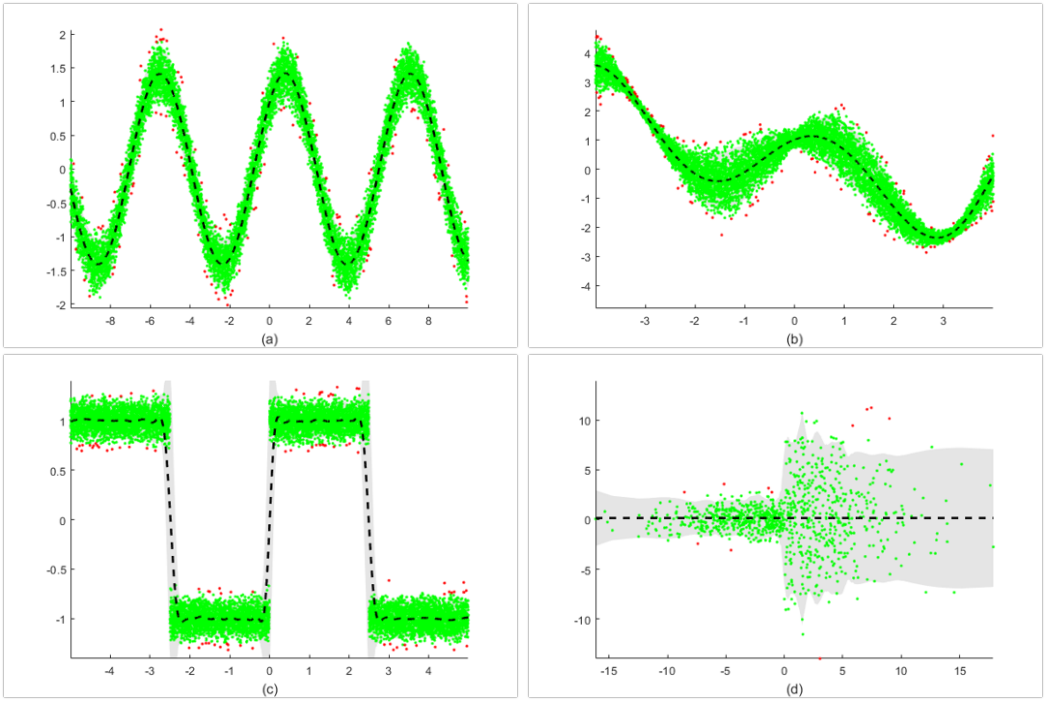


Figure B.4: Examples of the Gaussian process regression. Panel (a) and (c) are the homoscedastic cases whereas (b) and (d) are for the heteroscedastic data.

Then, one can show that the posterior $p(\beta|\mathcal{X},\mathcal{Y})$ is analytically given as follows:

$$\begin{aligned} p(\beta|\mathcal{X},\mathcal{Y}) &= \mathcal{N}\left(\beta|\bar{\mu}(\mathcal{X},\mathcal{Y}),\bar{\Sigma}(\mathcal{X},\mathcal{Y})\right) \\ \bar{\mu}(\mathcal{X},\mathcal{Y}) &= \mu_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}\right) \\ \bar{\Sigma}(\mathcal{X},\mathcal{Y}) &= \mathbb{K}_{\mathcal{X}\mathcal{X}} - \mathbb{K}_{\mathcal{X}\mathcal{X}}\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\mathbb{K}_{\mathcal{X}\mathcal{X}} \end{aligned} \tag{B.26}$$

By plugging (B.26) together with (B.25) and (B.2) into (B.1), we have the following posterior predictive:

$$\begin{aligned} p(y|x,\mathcal{X},\mathcal{Y}) &= \mathcal{N}\left(y|\hat{\mu}(x),\Lambda_x + \hat{\nu}(x)\right) \\ \hat{\mu}(x) &= \mu_x + \mathbb{K}_{x\mathcal{X}}\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}\right) \\ \hat{\nu}(x) &= \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}}\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\mathbb{K}_{\mathcal{X}x} \end{aligned} \tag{B.27}$$

In other words, a GPR returns the posterior predictive distribution as the regression model in a closed form: we will revisit this form later in section B.7 that discusses the actual tractability of (B.27) in practice.

Let us look deeper into (B.27). The variance term $\Lambda_x + \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}} \left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1} \mathbb{K}_{\mathcal{X}x}$ is decomposed into Λ_x and $\mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}} \left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1} \mathbb{K}_{\mathcal{X}x}$. Λ_x comes from the observational uncertainty in (B.25) while the rest of them stem from the model uncertainty obtained by marginalizing the latent states out: because the latter varies over x unless the kernel is constant (B.4), all GPR is heteroscedastic over inputs in principle. However, the resulting model is often suboptimal if the heteroscedasticity is let depend on $\mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}} \left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1} \mathbb{K}_{\mathcal{X}x}$ only, thus one might consider the observational heteroscedasticity as well. Homoscedastic GPR (section B.2.1) assumes that Λ is given as a constant function; heteroscedastic GPR (section B.2.2), however, regards Λ as a function that could change over the inputs. Figure B.4 gives some examples of the GPR.

There is one golden rule in modeling based on the Gaussian distributions including GP: they are very sensitive to the outliers! In the real data, it is often the case that some of them are just noninformative noises. Because GP is a nonparametric method that depends directly on the data, classifying (and discarding) outliers is indispensable: see [Kaiser et al. (2018); Lee & Lawrence (2019)] for more details. Another way of dealing with outliers is to take non-Gaussian observational models that will be discussed in section B.2.3.

B.2.1 Homoscedastic Gaussian Process Regression

A homoscedastic GPR assumes that $\Lambda \equiv \lambda_0^2 \mathbb{I}$ for some constant $\lambda_0 > 0$. Thus, the previous formulations are rewritten as follows:

1. Observational model:

$$\begin{aligned} p\left(\mathcal{Y}|\mathcal{X}, \beta\right) &= \mathcal{N}\left(\mathcal{Y}|\beta, \lambda_0^2 \mathbb{I}\right) = \prod_{n=1}^{N} \mathcal{N}\left(\mathrm{Y}_n|\beta_n, \lambda_0^2\right) \\ p\left(y|x, b\right) &= \mathcal{N}\left(y|b, \lambda_0^2\right) \end{aligned} \tag{B.28}$$

2. Posterior distribution of the latent state:

$$\begin{aligned} p\left(\beta|\mathcal{X},\mathcal{Y}\right) &= \mathcal{N}\left(\beta\middle|\overline{\mu}\left(\mathcal{X},\mathcal{Y}\right),\overline{\Sigma}\left(\mathcal{X},\mathcal{Y}\right)\right) \\ \overline{\mu}\left(\mathcal{X},\mathcal{Y}\right) &= \mu_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\left(\lambda_0^2\mathbb{I} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}\right) \\ \overline{\Sigma}\left(\mathcal{X},\mathcal{Y}\right) &= \mathbb{K}_{\mathcal{X}\mathcal{X}} - \mathbb{K}_{\mathcal{X}\mathcal{X}}\left(\lambda_0^2\mathbb{I} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\mathbb{K}_{\mathcal{X}\mathcal{X}} \end{aligned} \tag{B.29}$$

3. Posterior predictive as a regression model:

$$\begin{aligned} p\left(y|x,\mathcal{X},\mathcal{Y}\right) &= \mathcal{N}\left(y\middle|\hat{\mu}\left(x\right),\lambda_0^2 + \hat{\nu}\left(x\right)\right) \\ \hat{\mu}\left(x\right) &= \mu_x + \mathbb{K}_{x\mathcal{X}}\left(\lambda_0^2\mathbb{I} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}\right) \\ \hat{\nu}\left(x\right) &= \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}}\left(\lambda_0^2\mathbb{I} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)^{-1}\mathbb{K}_{\mathcal{X}x} \end{aligned} \tag{B.30}$$

Here, the objective function to be maximized for tuning the kernel hyperparameters is given as the following ML:

$$p\left(\mathcal{Y}|\mathcal{X}\right) = \int p\left(\mathcal{Y}|\mathcal{X},\beta\right)p\left(\beta|\mathcal{X}\right)d\beta = \mathcal{N}\left(\mathcal{Y}\middle|\mu_{\mathcal{X}},\lambda_0^2\mathbb{I} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right) \tag{B.31}$$

Under the assumption that the mean function μ has already been identified, the hyperparameters to tune are λ_0 and those in the kernel $\mathbb{K}$. Gradient descent algorithms on the log-ML are considered for the optimization, or, if only a few and low-dimensional hyperparameters are given, picking the most probable tuple among the finite candidates could be an alternative.

The mean prior function μ reflects a prior belief on the hidden state (the regression here). It could be given as a constant vector (including $\vec{0}$) or parametric function that had been either specified (with known parameters) or not (with parameters to learn). Even if μ is a parametric function to be learned, it can be done by maximizing (B.31).

The leave-one-out cross-validation (LOO-CV) is an alternative to define the tuning procedure. Instead of maximizing the ML in (B.31), it attempts to optimize the LOO log predictive likeli-

hood $\mathcal{L}_{\mathrm{LOO}}$ as follows: details can be found in [Rasmussen & Williams (2006)].

$$\mathcal{L}_{\mathrm{LOO}} = \sum_{n=1}^{\mathrm{N}} \log p\left(\mathrm{Y}_n | \mathcal{X}, \mathcal{Y} \setminus \mathrm{Y}_n\right) \tag{B.32}$$

The ML uses the likelihood of the observations under the assumption that the GP model is correctly specified. Thus, there is nothing strange about the argument that LOO-CV is more robust against model misspecification [Wahba (1990)].

B.2.2 Heteroscedastic Gaussian Process Regression

Contrary to the homoscedastic GPR, a heteroscedastic GPR assumes the general variance function Λ. If Λ is given a priori, then what remains is to tune kernel hyperparameters with the following objective function to maximize:

$$p\left(\mathcal{Y}|\mathcal{X}\right) = \int p\left(\mathcal{Y}|\mathcal{X}, \beta\right) p\left(\beta|\mathcal{X}\right) d\beta = \mathcal{N}\left(\mathcal{Y}|\mu_{\mathcal{X}}, \Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{X}}\right) \tag{B.33}$$

If Λ is also supposed to be learned, however, the problem becomes more complicated: first of all, there is no target variance, unlike the observations to be regressed are given. Defining Λ to be parametric and learning those parameters by maximizing (B.33) would be an answer, but it could be inflexible or misspecified.

Though there is no perfect answer that addresses this problem, a lot of relevant literatures have been published. [Goldberg et al. (1998); Le et al. (2005)] consider the noise itself to be a random variable and estimate it along with the mean of the distribution. [Kersting et al. (2007); Wang & Chen (2016)] independently infer Λ from the sampled variances given the regression model, step-by-step. [Lázaro-Gredilla & Titsias (2011)] parametrizes Λ as the exponential function where the exponent follows a GP prior. [Binois et al. (2018)] assumes a latent variable for constructing an apparatus to jointly infer a spatial field of simulation variances along with the mean-field GP.

[Lee & Lawrence (2019)] focuses on the structure of the observational model: $p\left(y|x, b\right) = \mathcal{N}\left(y|b, \Lambda_x\right)$ implies that the variance function Λ_x is supposed to be the variance of y given b as

the mean, which is equivalent to the expectation of $(y-b)^2$. Though y and b are not given at an arbitrary query input x, $(y-b)^2$ can be regressed given $(\mathcal{X}, \beta, \mathcal{Y})$ by the following Nadaraya-Watson kernel regression [Nadaraya (1964); Watson (1964); Bierens (1994); Langrené & Warin (2019)]:

$$(y-b)^2 \approx \frac{\sum_{n=1}^{\mathrm{N}} \left(\mathrm{Y}_n - \beta_n\right)^2 \mathcal{K}_h\left(x - \mathrm{X}_n\right)}{\sum_{n=1}^{\mathrm{N}} \mathcal{K}_h\left(x - \mathrm{X}_n\right)} \tag{B.34}$$

, where $\mathcal{K}$ and h are the density kernel and bandwidth parameter, respectively.

One advantage of the GP regression model is that the posterior $p\left(\beta|\mathcal{X}, \mathcal{Y}\right)$ is explicitly given. Thus, one can compute the following expectation analytically:

$$\begin{aligned} \Lambda_x &\triangleq \mathbb{E}_{\beta|\mathcal{X},\mathcal{Y}}\left[\frac{\sum_{n=1}^{\mathrm{N}} \left(\mathrm{Y}_n - \beta_n\right)^2 \mathcal{K}_h\left(x - \mathrm{X}_n\right)}{\sum_{n=1}^{\mathrm{N}} \mathcal{K}_h\left(x - \mathrm{X}_n\right)}\right] \\ &= \frac{\sum_{n=1}^{\mathrm{N}} \left(\left(\mathrm{Y}_n - \hat{\mu}\left(\mathrm{X}_n\right)\right)^2 + \hat{\nu}\left(\mathrm{X}_n\right)\right) \mathcal{K}_h\left(x - \mathrm{X}_n\right)}{\sum_{n=1}^{\mathrm{N}} \mathcal{K}_h\left(x - \mathrm{X}_n\right)} \end{aligned} \tag{B.35}$$

The subsequent problem is that computing each Λ_x requires the values of $\Lambda_{\mathcal{X}}$ at $\mathcal{X} = \{\mathrm{X}_n\}_{n=1}^{\mathrm{N}}$ a priori in (B.35). This can be addressed by iterating (B.35) to update $\Lambda_{\mathcal{X}}$ until convergence. The performance of the heteroscedastic GPR is substantially dependent on the bandwidth h: one suggestion is the K-nearest neighbor bandwidth [Loftsgaarden & Quesenberry (1965); Terrell & Scott (1992)] with the LOO-CV for choosing K.

As mentioned at the beginning of the section, tuning hyperparameters is straightforward once $\Lambda_{\mathcal{X}}$ is specified. Thus, the heteroscedastic GP regression can be done by iterating the following steps:

1. Choose h and K given $\mathbb{K}$ and $\Lambda_{\mathcal{X}}$.
2. Update $\Lambda_{\mathcal{X}}$ given h, K and $\mathbb{K}$.
3. Tune $\mathbb{K}$ given $\Lambda_{\mathcal{X}}$.

B.2.3 Non-Gaussian Observational Models

So far, observational models have been assumed to be Gaussian. As discussed earlier in this section, it could be too strong to assume that the standardized residuals $\mathcal{R} \triangleq \{r_n\}_{n=1}^{\mathrm{N}}$ would follow the standard normal distribution in practice, where:

$$r_n \triangleq \frac{\mathrm{Y}_n - \hat{\mu}\left(\mathrm{X}_n\right)}{\sqrt{\Lambda_{\mathrm{X}_n} + \hat{\nu}\left(\mathrm{X}_n\right)}} \tag{B.36}$$

The model misspecification could be from the multimodality of the data [Lázaro-Gredilla et al. (2012); Bodin et al. (2017); Kaiser et al. (2018)], but here we just focus on the misspecification of the observational models.

Let us go back to the posterior predictive in (B.1): $p\left(y|x, \mathcal{X}, \mathcal{Y}\right)$ is expressed as a double integration over the hidden states β and b. If the observational model is assumed to be Gaussian, then not only $p\left(y|x, b\right)$ but also $p\left(\beta|\mathcal{X}, \mathcal{Y}\right)$ are also Gaussian, thus computing (B.1) is done explicitly. If not, it is even not guaranteed to get $p\left(\beta|\mathcal{X}, \mathcal{Y}\right)$ as a known distribution. One of the next best things is to sample a number of β from $p\left(\beta|\mathcal{X}, \mathcal{Y}\right) \propto p\left(\beta|\mathcal{X}\right) p\left(\mathcal{Y}|\beta, \mathcal{X}\right)$ and approximate (B.1) as follows: if $p\left(\mathcal{Y}|\beta, \mathcal{X}\right)$ is differentiable with respect to β, then the Hamiltonian Monte Carlo (HMC) algorithm in section A.4.3 can be applied for the sampling. Let $\left\{\widetilde{\beta}^{(m)}\right\}_{m=1}^{\mathrm{M}}$ be such samples, then:

$$p\left(y|x, \mathcal{X}, \mathcal{Y}\right) \approx \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \int p\left(y|x, b\right) \mathcal{N}\left(b \middle| \mu_x + \mathbb{K}_{x\mathcal{X}} \mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1} \left(\widetilde{\beta}^{(m)} - \mu_{\mathcal{X}}\right), \hat{\nu}\left(x\right)\right) db \tag{B.37}$$

The other approach corresponds to the approximation of $p\left(\beta|\mathcal{X}, \mathcal{Y}\right)$ with a known distribution, $q\left(\beta|\theta\right)$, where θ is a variational parameter. Similar to the discussion in A.1.4, we might consider the following Kullback-Leibler (KL) divergence:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{X}, \mathcal{Y}\right)\right) \\ &= \log p\left(\mathcal{Y}|\mathcal{X}\right) + \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{X}\right)\right) - \mathbb{E}_{q(\beta|\theta)}\left[\log p\left(\mathcal{Y}|\beta, \mathcal{X}\right)\right] \geq 0 \end{aligned} \tag{B.38}$$

Therefore, the log-ML $\log p\left(\mathcal{Y}|\mathcal{X}\right)$ has the ELBO $\mathcal{L}$ as follows:

$$\log p\left(\mathcal{Y}|\mathcal{X}\right) \geq \mathbb{E}_{q(\beta|\theta)}\left[\log p\left(\mathcal{Y}|\beta,\mathcal{X}\right)\right] - \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{X}\right)\right) = \mathcal{L} \tag{B.39}$$

Let us assume further that $q\left(\beta|\theta\right)$ is given a fully independent Gaussian distribution:

$$q\left(\beta|\theta\right) \triangleq \mathcal{N}\left(\beta|\ddot{\mu}, diag\left(\ddot{\sigma}^2\right)\right) = \prod_{n=1}^{\mathrm{N}} \mathcal{N}\left(\beta_n|\ddot{\mu}_n, \ddot{\sigma}_n^2\right) \tag{B.40}$$

Then, the second term $\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{X}\right)\right)$ in (B.39) is explicitly given as follows:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{X}\right)\right) \\ &= \frac{1}{2}\left(\sum_{n=1}^{\mathrm{N}} \frac{\ddot{\sigma}_n^2}{\mathbb{K}_{\mathrm{X}_n\mathrm{X}_n}} + \left(\mu_{\mathcal{X}} - \ddot{\mu}\right)^{\mathbb{T}} \mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\left(\mu_{\mathcal{X}} - \ddot{\mu}\right) - 1 + \log\left|\mathbb{K}_{\mathcal{X}\mathcal{X}}\right| - \sum_{n=1}^{\mathrm{N}} \log \ddot{\sigma}_n^2\right) \end{aligned} \tag{B.41}$$

The first term is decomposed as follows:

$$\mathbb{E}_{q(\beta|\theta)}\left[\log p\left(\mathcal{Y}|\beta,\mathcal{X}\right)\right] = \sum_{n=1}^{\mathrm{N}} \mathbb{E}_{\mathcal{N}\left(\beta_n|\ddot{\mu}_n,\ddot{\sigma}_n^2\right)}\left[\log p\left(\mathrm{Y}_n|\beta_n,\mathrm{X}_n\right)\right] \tag{B.42}$$

The reparameterization trick [Kingma & Welling (2013)] gives a way of reducing variances in computing the numerical gradient of $\mathbb{E}_{q(\beta|\theta)}\left[\log p\left(\mathcal{Y}|\beta,\mathcal{X}\right)\right]$ with respect to the variational parameters:

$$\nabla_\theta \mathbb{E}_{q(\beta|\theta)}\left[\log p\left(\mathcal{Y}|\beta,\mathcal{X}\right)\right] = \sum_{n=1}^{\mathrm{N}} \mathbb{E}_{\mathcal{N}(\varepsilon|0,1)}\left[\nabla_\theta \log p\left(\mathrm{Y}_n|\ddot{\mu}_n + \ddot{\sigma}_n\varepsilon, \mathrm{X}_n\right)\right] \tag{B.43}$$

Once learning/tuning the variational parameters/kernel hyperparameters with $\mathcal{L}$ as the objective function are done, we can approximate $p\left(y|x,\mathcal{X},\mathcal{Y}\right)$ by replacing $p\left(\beta|\mathcal{X},\mathcal{Y}\right)$ with $q\left(\beta|\theta\right)$ in (B.1):

$$\begin{aligned} p\left(y|x,\mathcal{X},\mathcal{Y}\right) &\approx \int p\left(y|x,b\right) \int p\left(b|\beta,x,\mathcal{X}\right) q\left(\beta|\theta\right) d\beta db \\ &= \int p\left(y|x,b\right) \mathcal{N}\left(b|\widetilde{\mu}\left(x\right), \widetilde{\nu}\left(x\right)\right) db \end{aligned} \tag{B.44}$$

, where:

$$\begin{aligned}\widetilde{\mu}(x) &= \mu_x + \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\left(\ddot{\mu} - \mu_{\mathcal{X}}\right) \\ \widetilde{\nu}(x) &= \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}x} + \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}diag\left(\ddot{\sigma}^2\right)\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}x}\end{aligned} \tag{B.45}$$

Note that the above analytic computation is available if q is a (multivariate) Gaussian. (B.40) is just the simplest choice for the variational approximation: we will discuss further about the choice of q in section B.4.

B.3 Gaussian Process Classification

Classification is a function that maps each query input to one (or more) of the discrete categories (labels). Specifically, statistical classification gives a probabilistic model as the classification function. Gaussian process classification (GPC) [Rasmussen & Williams (2006); Barber (2012)] is a nonparametric classification method that assigns a discrete probability distribution to each query input, based on the data. Here we consider the binary classification only (in other words, each $Y_n \in \{-1, +1\}$), though multi-class extension is straightforward [Rasmussen & Williams (2006); Hernández-Lobato et al. (2011); Murphy (2012); Villacampa-Calvo et al. (2020)]. A graphical representation is given in figure B.3.

Contrary to the GPR models in section B.2, a GPC regards the latent state β as a predictor (log-odd) function (over $\mathcal{X}$) that is squashed through the logistic function, so $p(\mathcal{Y}|\mathcal{X},\beta) = p(\mathcal{Y}|\beta)$ in (B.1) is replaced with the following:

$$\begin{aligned}p(\mathcal{Y}|\beta) &= \prod_{n=1}^{N}\sigma\left(Y_n\beta_n\right) \\ p(y|b) &= \sigma(yb)\end{aligned} \tag{B.46}$$

, whereas we assume the same prior on β, $p(\beta|\mathcal{X}) = \mathcal{GP}\left(\beta\middle|\vec{0}, \mathbb{K}_{\mathcal{X}\mathcal{X}}\right)$ as usual with $\mu \equiv \vec{0}$ (note that β now represents a log-odd), and $\sigma : \mathbb{R} \to (0,1)$ is a logistic function that is also defined as follows:

$$\sigma(t) = \frac{1}{1+\exp(-t)} \tag{B.47}$$

Note that (B.46) that is non-Gaussian prevents from computing both posterior predictive $p(y|x,\mathcal{X},\mathcal{Y})$ and posterior $p(\beta|\mathcal{X},\mathcal{Y})$ analytically. Thus, it is inevitable to consider either Monte Carlo sampling or approximation for accessing $p(\beta|\mathcal{X},\mathcal{Y})$, as discussed in section B.2.3. In addition, characteristics of the classification bring other approximation methods such as Laplace approximation (section B.3.1) and expectation propagation (section B.3.2). Moreover, we will cover one of the most recent variational approach that uses the variable augmentation technique in section B.3.3.

B.3.1 Laplace Approximation

The Laplace approximation [Williams & Barber (1998)] aims at approximating the posterior $p(\beta|\mathcal{X},\mathcal{Y})$ that is non-Gaussian with a Gaussian distribution $q(\beta)$ that is obtained by doing a second-order Taylor expansion of $\log p(\beta|\mathcal{X},\mathcal{Y})$ around its maximizer $\hat{\beta}$:

$$q(\beta) = \mathcal{N}\left(\beta\middle|\hat{\beta}, \left(-\nabla\nabla \log p(\beta|\mathcal{X},\mathcal{Y})|_{\beta=\hat{\beta}}\right)^{-1}\right) \tag{B.48}$$

Because $\log p(\beta|\mathcal{X},\mathcal{Y}) = \log p(\mathcal{Y}|\beta) + \log p(\beta|\mathcal{X}) - \log p(\mathcal{Y}|\mathcal{X}) \triangleq \mathcal{L}(\beta) - \log p(\mathcal{Y}|\mathcal{X})$,

$$\begin{aligned} \hat{\beta} &= \underset{\beta}{\operatorname{argmax}} \log p(\beta|\mathcal{X},\mathcal{Y}) = \underset{\beta}{\operatorname{argmax}} \mathcal{L}(\beta) \\ \nabla\nabla \log p(\beta|\mathcal{X},\mathcal{Y}) &= \nabla\nabla \mathcal{L}(\beta) \end{aligned} \tag{B.49}$$

Since $\mathrm{Y}_n^2 \equiv 1$, we have:

$$\begin{aligned} \mathcal{L}(\beta) &= -\sum_{n=1}^{\mathrm{N}} \log\left(1+\exp\left(-\mathrm{Y}_n\beta_n\right)\right) - \frac{1}{2}\beta^{\mathrm{T}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\beta - \frac{1}{2}\log|\mathbb{K}_{\mathcal{X}\mathcal{X}}| \\ \nabla\mathcal{L}(\beta) &= \nabla \log p(\mathcal{Y}|\beta) - \mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\beta \\ \nabla\nabla\mathcal{L}(\beta) &= \nabla\nabla \log p(\mathcal{Y}|\beta) - \mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1} \end{aligned} \tag{B.50}$$

, where:

$$\begin{aligned}\nabla \log p\left(\mathcal{Y}|\beta\right) &= \begin{pmatrix} \mathrm{Y}_1\left(1+\exp\left(\mathrm{Y}_1\beta_1\right)\right)^{-1} \\ \mathrm{Y}_2\left(1+\exp\left(\mathrm{Y}_2\beta_2\right)\right)^{-1} \\ \vdots \\ \mathrm{Y}_\mathrm{N}\left(1+\exp\left(\mathrm{Y}_\mathrm{N}\beta_\mathrm{N}\right)\right)^{-1} \end{pmatrix} \\ \nabla\nabla \log p\left(\mathcal{Y}|\beta\right) &= diag\left(\begin{pmatrix} -\sigma\left(\mathrm{Y}_1\beta_1\right)\left(1-\sigma\left(\mathrm{Y}_1\beta_1\right)\right) \\ -\sigma\left(\mathrm{Y}_2\beta_2\right)\left(1-\sigma\left(\mathrm{Y}_2\beta_2\right)\right) \\ \vdots \\ -\sigma\left(\mathrm{Y}_\mathrm{N}\beta_\mathrm{N}\right)\left(1-\sigma\left(\mathrm{Y}_\mathrm{N}\beta_\mathrm{N}\right)\right) \end{pmatrix}\right)\end{aligned} \tag{B.51}$$

Finding $\hat{\beta}$ can be done numerically by the Newton's method, i.e., iterating the following until convergence:

$$\begin{aligned}&\beta^{(t+1)} \triangleq \beta^{(t)} - \left(\nabla\nabla\mathcal{L}\left(\beta^{(t)}\right)\right)^{-1}\left(\nabla\mathcal{L}\left(\beta^{(t)}\right)\right) \\ &= \beta^{(t)} + \left(\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1} - \nabla\nabla \log p\left(\mathcal{Y}|\beta\right)|_{\beta=\beta^{(t)}}\right)^{-1}\left(\nabla \log p\left(\mathcal{Y}|\beta\right)|_{\beta=\beta^{(t)}} - \mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\beta^{(t)}\right)\end{aligned} \tag{B.52}$$

Once $\hat{\beta}$ is obtained, $q\left(\beta\right) = \mathcal{N}\left(\beta\middle|\hat{\beta}, \left(\nabla\nabla\mathcal{L}\left(\hat{\beta}\right)\right)^{-1}\right)$ is derived from (B.48) and (B.50), and the posterior predictive distribution as the classifier is approximated as follows:

$$\begin{aligned}p\left(y|x,\mathcal{X},\mathcal{Y}\right) &\approx \int p\left(y|b\right)\int p\left(b|\beta,x,\mathcal{X}\right)q\left(\beta\right)d\beta db \\ &= \int \sigma\left(yb\right)\mathcal{N}\left(b|\widetilde{\mu}\left(x\right),\widetilde{\nu}\left(x\right)\right)db\end{aligned} \tag{B.53}$$

, where:

$$\begin{aligned}&\widetilde{\mu}\left(x\right) = \mu_x + \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\left(\hat{\beta}-\mu_{\mathcal{X}}\right) \\ &\widetilde{\nu}\left(x\right) = \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}x} + \mathbb{K}_{x\mathcal{X}}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\left(\nabla\nabla\mathcal{L}\left(\hat{\beta}\right)\right)^{-1}\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}x}\end{aligned} \tag{B.54}$$

Tuning kernel hyperparameters can be done by maximizing the approximated marginal likelihood with $q\left(\beta\right)$: details are in [Barber (2012)].

B.3.2 Expectation Propagation

Just as the Laplace approximation in the previous subsection, the expectation propagation (EP) [Minka (2001)] method also seeks the approximation of $p(\beta|\mathcal{X},\mathcal{Y})$ with a Gaussian $q(\beta)$. The difference comes from the form of $q(\beta)$: it is designed to be proportional to the product of the GP prior $p(\beta|\mathcal{X})$ and pointwise inferences on β.

The core idea starts from approximating each $p\left(\mathrm{Y}_n|\beta_n\right)$ with $\mathcal{N}\left(\beta_n|\widetilde{\mu}_n,\widetilde{\nu}_n\right)$ up to constant product of $\widetilde{\mathrm{Z}}_n$ as follows:

$$p\left(\mathrm{Y}_n|\beta_n\right) \approx \widetilde{\mathrm{Z}}_n \mathcal{N}\left(\beta_n|\widetilde{\mu}_n,\widetilde{\nu}_n\right) \tag{B.55}$$

Then define $q(\beta)$ from (B.55) and $p(\beta|\mathcal{X})$ as follows:

$$\begin{aligned} q(\beta) &= \mathcal{N}\left(\beta\middle|\left(\mathbb{I}+diag\left(\widetilde{\nu}\right)\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\right)^{-1}\widetilde{\mu},\left(\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}+diag\left(\widetilde{\nu}\right)^{-1}\right)^{-1}\right) \\ &\propto p(\beta|\mathcal{X})\prod_{n=1}^{\mathrm{N}}\mathcal{N}\left(\beta_n|\widetilde{\mu}_n,\widetilde{\nu}_n\right) \end{aligned} \tag{B.56}$$

The inference on $\widetilde{\mu}$ and $\widetilde{\nu}$ is done by sequential optimization: details can be found in [Rasmussen & Williams (2006)]. Note that defining $p(\mathcal{Y}|\beta)$ with the probit likelihood $\Phi(t) = \int_{-\infty}^{t}\mathcal{N}(x|0,1)\,dx$ gives not only the explicit updating form for $\widetilde{\mu}$ and $\widetilde{\nu}$ but also the explicit posterior predictive approximation (that is advantageous over the Laplace approximation) as follows:

$$p(y=+1|x,\mathcal{X},\mathcal{Y}) \approx \Phi\left(\frac{\mathbb{K}_{x\mathcal{X}}\left(\mathbb{K}_{\mathcal{X}\mathcal{X}}+diag\left(\widetilde{\nu}\right)\right)^{-1}\widetilde{\mu}}{\sqrt{1+\mathbb{K}_{xx}-\mathbb{K}_{x\mathcal{X}}\left(\mathbb{K}_{\mathcal{X}\mathcal{X}}+diag\left(\widetilde{\nu}\right)\right)^{-1}\mathbb{K}_{\mathcal{X}x}}}\right) \tag{B.57}$$

B.3.3 Polya-Gamma Data Augmentation

The Polya-Gamma data augmentation [Wenzel et al. (2018)] begins with exploiting the following property of the logistic function:

$$\sigma(z) = \frac{1}{1+\exp(-z)} = \frac{1}{2}\int \exp\left(\frac{z}{2}-\frac{z^2}{2}w\right)\mathrm{PG}(w|1,0)\,dw \tag{B.58}$$

, where $\mathrm{PG}(w|1,c)$ is the Polya-Gamma distribution defined by the following moment generating function and proportionality:

$$\begin{aligned}\mathbb{E}_{\mathrm{PG}(w|1,0)}\left[\exp\left(-wt\right)\right] &= \frac{1}{\cosh\left(\sqrt{t/2}\right)} \\ \mathrm{PG}\left(w|1,c\right) &\propto \exp\left(-\frac{c^2}{2}w\right)\mathrm{PG}\left(w|1,0\right)\end{aligned} \tag{B.59}$$

From (B.58), we have the following proportionality with an augmented state $\mathcal{W} = \mathrm{W}_{1:\mathrm{N}}$:

$$p\left(\mathcal{Y},\mathcal{W}|\beta,\mathcal{X}\right) \propto \exp\left(-\frac{1}{2}\mathcal{Y}^{\mathbb{T}}\beta - \frac{1}{2}\beta^{\mathbb{T}}diag\left(\mathcal{W}\right)\beta\right)\prod_{n=1}^{\mathrm{N}}\mathrm{PG}\left(\mathrm{W}_n|1,0\right) \tag{B.60}$$

The key idea is to approximate $p\left(\beta,\mathcal{W}|\mathcal{X},\mathcal{Y}\right) \propto p\left(\mathcal{Y},\mathcal{W}|\beta,\mathcal{X}\right)p\left(\beta|\mathcal{X}\right)$ with the following $q\left(\beta,\mathcal{W}|\mu,\Sigma,\mathcal{C}\right)$ with variational parameters μ, Σ and $\mathcal{C} = c_{1:\mathrm{N}}$:

$$q\left(\beta,\mathcal{W}|\mu,\Sigma,\mathcal{C}\right) \triangleq \mathcal{N}\left(\beta|\mu,\Sigma\right)\prod_{n=1}^{\mathrm{N}}\mathrm{PG}\left(\mathrm{W}_n|1,c_n\right) \tag{B.61}$$

The KL divergence $\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\mu,\Sigma,\mathcal{C}\right)||p\left(\cdot|\mathcal{X},\mathcal{Y}\right)\right)$ results in the following ELBO $\mathcal{L}$:

$$\begin{aligned}&\log p\left(\mathcal{Y}|\mathcal{X}\right) \geq \mathcal{L} \\ &\triangleq \mathbb{E}_{q(\beta,\mathcal{W}|\mu,\Sigma,\mathcal{C})}\left[\log p\left(\mathcal{Y}|\beta,\mathcal{W},\mathcal{X}\right)\right] - \mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\mu,\Sigma,\mathcal{C}\right)\middle\|p\left(\cdot|\mathcal{X}\right)\prod_{n=1}^{\mathrm{N}}\mathrm{PG}\left(\mathrm{W}_n|1,0\right)\right)\end{aligned} \tag{B.62}$$

The variational parameters are learned by maximizing $\mathcal{L}$. One advantage of this approach is that $\mathcal{L}$ is given in a closed form and $\mathcal{C} = c_{1:\mathrm{N}}$ can be updated analytically: see [Wenzel et al. (2018)] for more details. Once variational parameters μ, Σ and $\mathcal{C}$ are learned, (B.61) gives the marginal $q\left(\beta|\mu,\Sigma,\mathcal{C}\right) = \mathcal{N}\left(\beta|\mu,\Sigma\right)$ directly. What remains is to approximate the posterior predictive by replacing $p\left(\beta|\mathcal{X},\mathcal{Y}\right)$ with q, as done in section B.3.1.

Figure B.5 presents two examples of the GP classification based on the Polya-Gamma data augmentation. The left panel uses 10,000 observations whereas the right one has 1,000. Note

that the right example results in less certain posteriors than the left one: this suggests an advantage of the nonparametric classification algorithm that the inference directly depends on the quality, quantity and sparsity of the data.

B.4 Gaussian Process State-space Models

State-space models in chapter A assume the Markov property and well-established transition models. Those are too strong assumptions to cover all real problems: such a 'memoryless' property has potential of losing too many information from the previously observed states and observations would not be made regularly. Recurrent neural networks (RNNs) [Goodfellow et al. (2016)] straightforwardly address the first case, however, usually require too many parameters thus would not be appropriate if only a small number of observations are given.

Gaussian process state-space models (GPSTs) [Frigola et al. (2014); Eleftheriadis et al. (2017)] are good alternatives to those models that depend on the Markov assumption. It gives a GP prior on the series of each predictor: a typical graphical representation is described in figure B.6. In this section, we follow the same notations in chapter A. Suppose that the hidden states $\mathcal{X} = \{\mathrm{X}_n\}_{n=1}^{\mathrm{N}}$ are associated with the pair of indices and outputs $(\mathcal{T}, \mathcal{Y})$, and each $\mathrm{X}_n \in \mathbb{R}^{\mathrm{D}}$ for $\mathrm{D} \in \mathbb{N}$. Then we could consider $\mathcal{X}$ as a $\mathrm{N} \times \mathrm{D}$ matrix where each row indicates X_n and the following is assumed for each $d = 1, 2, \cdots, \mathrm{D}$:

$$p\left(\mathrm{X}_{:,d} | \mathcal{T}\right) = \mathcal{GP}\left(\mathrm{X}_{:,d} \middle| \mu_{\mathcal{T}}^{(d)}, \mathbb{K}_{\mathcal{T}\mathcal{T}}^{(d)}\right) \tag{B.63}$$

Just as the other state-space models in chapter A, $\mathcal{Y}$ is also defined on a high-dimensional space $\mathbb{R}^{\mathrm{L}}$ and the emission model is given as $p_e\left(\mathrm{Y}_n | \mathrm{T}_n, \mathrm{X}_n\right)$ for each n: now the possible nonstationary characteristic is controlled by the associated index (time in common) T_n.

Thus, the prior (B.64) and likelihood (B.65) are given as follows:

$$p\left(\mathcal{X} | \mathcal{T}\right) = \prod_{d=1}^{\mathrm{D}} p\left(\mathrm{X}_{:,d} | \mathcal{T}\right) = \prod_{d=1}^{\mathrm{D}} \mathcal{GP}\left(\mathrm{X}_{:,d} \middle| \mu_{\mathcal{T}}^{(d)}, \mathbb{K}_{\mathcal{T}\mathcal{T}}^{(d)}\right) \tag{B.64}$$

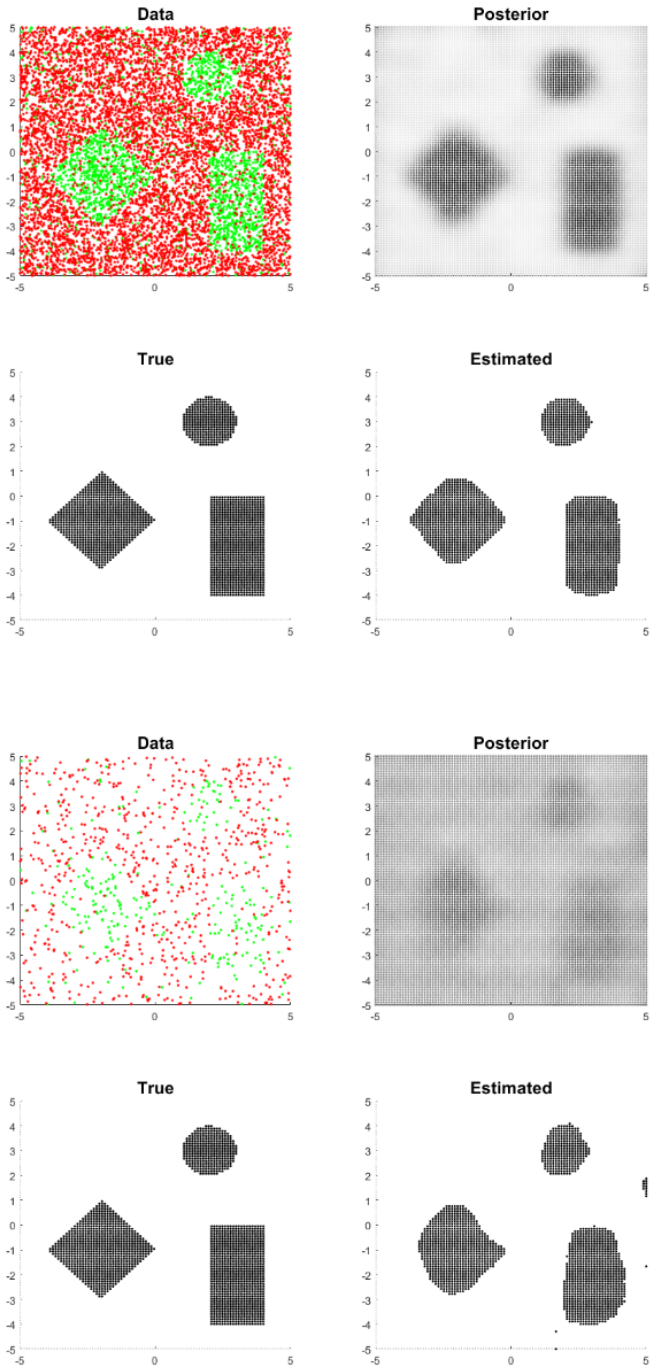


Figure B.5: Examples of the GP classification. Each panel consists of four sub-panels. The first sub-panel show the artificial data that are drawn from the distribution depicted in the third sub-panel with 10% contamination. The second and fourth sub-panels show the posterior distributions of the labels with representing their levels with the brightness and the corresponding estimation, respectively.

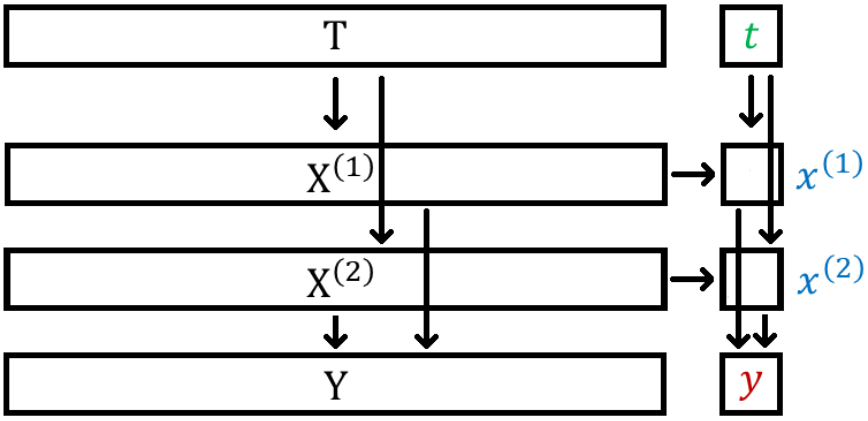


Figure B.6: A graphical representation of GPST. Capital T and Y are the training indices (time) and outputs, respectively, and $\mathrm{X}^{(1)}$ and $\mathrm{X}^{(2)}$ indicate the latent states, that are hidden states to infer as those of the state-space models in chapter A. Green t and red y are the query index (time) and the associated output, respectively, and blue $x^{(1)}$ and $x^{(2)}$ stand for the associated latent variables.

$$p\left(\mathcal{Y}|\mathcal{T},\mathcal{X}\right)=\prod_{n=1}^{\mathrm{N}}p_e\left(\mathrm{Y}_n|\mathrm{T}_n,\mathrm{X}_n\right) \tag{B.65}$$

From above, the posterior of $\mathcal{X}$ is given as follows:

$$p\left(\mathcal{X}|\mathcal{T},\mathcal{Y}\right)\propto p\left(\mathcal{X}|\mathcal{T}\right)p\left(\mathcal{Y}|\mathcal{T},\mathcal{X}\right) \tag{B.66}$$

In state-space models, the major concern is to compute the prediction of the hidden state x, $p\left(x|t,\mathcal{T},\mathcal{Y}\right)$, at an arbitrary query index t.

$$p\left(x|t,\mathcal{T},\mathcal{X},\mathcal{Y}\right)=\int p\left(x|t,\mathcal{T},\mathcal{X}\right)p\left(\mathcal{X}|\mathcal{T},\mathcal{Y}\right)d\mathcal{X} \tag{B.67}$$

, where $p\left(x|t,\mathcal{T},\mathcal{X}\right)$ is derived from the assumption (B.64):

$$\begin{aligned}&p\left(x|t,\mathcal{T},\mathcal{X}\right)\\&=\prod_{d=1}^{\mathrm{D}}\mathcal{N}\left(x_d\middle|\mu_t^{(d)}+\mathbb{K}_{t\mathcal{T}}^{(d)}\left(\mathbb{K}_{\mathcal{T}\mathcal{T}}^{(d)}\right)^{-1}\left(\mathrm{X}_{:,d}-\mu_{\mathcal{T}}^{(d)}\right),\mathbb{K}_{tt}^{(d)}-\mathbb{K}_{t\mathcal{T}}^{(d)}\left(\mathbb{K}_{\mathcal{T}\mathcal{T}}^{(d)}\right)^{-1}\mathbb{K}_{\mathcal{T}t}^{(d)}\right)\end{aligned} \tag{B.68}$$

The choice of kernel function $\mathbb{K}$ characterizes the underlying dynamics of the hidden states. As discussed in section B.1.1, Matérn kernels correspond to the continuous version of AR models. This property is quite useful. Suppose that one wants to design a state-space model for a given time series under the assumption that the series of hidden states follows AR (2) and in addition the indices in the time series are not regular: it is quite common in paleoceanography and paleoclimatology, as discussed in section 1.1.3. Rearrangement of the time series by interpolation would be required to fit the data into the parametric framework, but the GPST with a kernel $\mathbb{K}_{\text{Matern}}^{(3/2)}$ is working properly regardless of the index regularity. Moreover, it does not need the Markovian assumption whereas parametric state-space models in chapter A are intractable without it. Lastly, GPST is nonparametric thus can be called data-driven AR model, i.e., we can easily solve the problem that would be picky in the parametric setting by moving it to the nonparametric world.

In most of the cases, however, $p\left(\mathcal{X}|\mathcal{T},\mathcal{Y}\right)$ is not given in a closed form. The variational inference discussed in section A.1.4 is still valid in this case: infer the tractable approximation $q\left(\mathcal{X}|\theta\right)$ together with kernel hyperparameters that maximize the ELBO $\mathcal{L}$:

$$\log p\left(\mathcal{Y}|\mathcal{T}\right) \geq \mathbb{E}_{q(\mathcal{X}|\theta)}\left[\log p\left(\mathcal{Y}|\mathcal{T},\mathcal{X}\right)\right] - \mathbb{D}_{\text{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{T}\right)\right) \tag{B.69}$$

Once q is learned, $p\left(x|t,\mathcal{T},\mathcal{X},\mathcal{Y}\right)$ in (B.67) is approximated by following:

$$p\left(x|t,\mathcal{T},\mathcal{X},\mathcal{Y}\right) \approx \int p\left(x|t,\mathcal{T},\mathcal{X}\right) q\left(\mathcal{X}|\theta\right) d\mathcal{X} \tag{B.70}$$

This section ends with the discussion about the form of $q\left(\mathcal{X}|\theta\right)$. Because $\mathcal{X}$ is not a vector here, roughly we have the following ways of defining it:

1. Markov chain: it is good for sampling paths.

$$q\left(\mathcal{X}|\theta\right) = q\left(\mathrm{X}_1|\theta\right) \prod_{n=2}^{\mathrm{N}} \pi\left(\mathrm{X}_n|\mathrm{X}_{n-1},\theta_n\right) \tag{B.71}$$

2. Predictor-wise model: we will see this structure again in section B.6.

$$q\left(\mathcal{X}|\theta\right) = \prod_{d=1}^{\mathrm{D}} q\left(\mathrm{X}_{:,d}\middle|\theta^{(d)}\right) \tag{B.72}$$

3. The full Gaussian model: not practical if $p\left(\mathcal{X}|\mathcal{T},\mathcal{Y}\right)$ is multimodal.

$$q\left(\mathcal{X}|\theta\right) = \mathcal{N}\left(\begin{pmatrix}\mathrm{X}_{:,1}\\ \vdots \\ \mathrm{X}_{:,\mathrm{D}}\end{pmatrix}\middle|\mu,\Sigma\right) \tag{B.73}$$

Recently, [Eleftheriadis et al. (2017)] considers the RNN models for modelling q. With the reparameterization trick, the gradient of $\mathcal{L}$ is efficiently approximated thus variational parameters can be learned numerically regardless of the parametrization. If q is Gaussian, then the second term $\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\theta\right)||p\left(\cdot|\mathcal{T}\right)\right)$ is expressed in a closed form.

B.5 Gaussian Process Latent Variable Models

Dimensionality reduction is essential for dealing with the curse of dimensionality in the preprocessing step before applying nonparametric models. The Gaussian process latent variable model (GPLVM) [Lawrence (2005); Damianou et al. (2016)] designs a nonparametric dimensionality reduction model based on GP: [Titsias & Lawrence (2010)] revisit the GPLVM to show that a variational inference framework can marginalize the input variables. More extensions can be found in [Li & Chen (2016)]. However, in this subsection we just cover the original model which aims at optimizing the latent inputs. The below formulations are referred to [Lawrence (2005); Murphy (2012)].

Unlike the settings so far, only outputs $\mathcal{Y} = \{\mathrm{Y}_n\}_{n=1}^{\mathrm{N}} \subseteq \mathbb{R}^{\mathrm{L}}$ are given here. The goal is to find $\mathcal{X} = \{\mathrm{X}_n\}_{n=1}^{\mathrm{N}} \subseteq \mathbb{R}^{\mathrm{D}}$ as the low dimension reduction of $\mathcal{Y}$ where $\mathrm{D} \ll \mathrm{L}$. Thus, it is an unsupervised learning. To be specific, we have the following relationship with a $\mathrm{L} \times \mathrm{D}$ matrix W

and a positive constant σ_0:

$$p\left(\mathrm{Y}_n \middle| \mathrm{X}_n, \mathrm{W}, \sigma_0^2\right) = \mathcal{N}\left(\mathrm{Y}_n \middle| \mathrm{W}\mathrm{X}_n, \sigma_0^2 \mathbb{I}\right) \tag{B.74}$$

The probabilistic principal component analysis (PPCA) [Tipping & Bishop (1999)] assumes a prior $p\left(\mathrm{X}_n\right) = \mathcal{N}\left(\mathrm{X}_n \middle| \vec{0}, \mathbb{I}\right)$ on X_n to get the marginal likelihood $p\left(\mathrm{Y}_n \middle| \mathrm{W}, \sigma_0^2\right)$ and infer W that maximizes the following likelihood:

$$\begin{aligned} p\left(\mathcal{Y} \middle| \mathrm{W}, \sigma_0^2\right) &= \prod_{n=1}^{\mathrm{N}} p\left(\mathrm{Y}_n \middle| \mathrm{W}, \sigma_0^2\right) \\ &= \prod_{n=1}^{\mathrm{N}} \int p\left(\mathrm{Y}_n \middle| \mathrm{X}_n, \mathrm{W}, \sigma_0^2\right) p\left(\mathrm{X}_n\right) d\mathrm{X}_n \\ &= \prod_{n=1}^{\mathrm{N}} \mathcal{N}\left(\mathrm{Y}_n \middle| \vec{0}, \sigma_0^2 \mathbb{I} + \mathrm{W}^{\mathbb{T}} \mathrm{W}\right) \end{aligned} \tag{B.75}$$

The GPLVM starts from the dual problem of PPCA, that is, to assume a prior $p\left(\mathrm{W}_{l,:}^{\mathbb{T}}\right) = \mathcal{N}\left(\mathrm{W}_{l,:}^{\mathbb{T}} \middle| \vec{0}, \alpha^{-1}\mathbb{I}\right)$ with a positive constant $\alpha > 0$ for each $l = 1, 2, \cdots, \mathrm{L}$ and obtain the marginal likelihood $p\left(\mathcal{Y} | \mathcal{X}, \sigma_0^2, \alpha\right)$ that has the following form: let us regard Y as a $\mathrm{N} \times \mathrm{L}$ matrix where each row indicates Y_n and X as a $\mathrm{N} \times \mathrm{D}$ matrix where each row indicates X_n.

$$\begin{aligned} p\left(\mathcal{Y} \middle| \mathcal{X}, \sigma_0^2, \alpha\right) &= \int \prod_{l=1}^{\mathrm{L}} p\left(\mathrm{W}_{l,:}^{\mathbb{T}}\right) \prod_{n=1}^{\mathrm{N}} p\left(\mathrm{Y}_n \middle| \mathrm{X}_n, \mathrm{W}, \sigma_0^2\right) d\mathrm{W} \\ &= \int \prod_{l=1}^{\mathrm{L}} \mathcal{N}\left(\mathrm{W}_{l,:}^{\mathbb{T}} \middle| \vec{0}, \alpha^{-1}\mathbb{I}\right) \prod_{n=1}^{\mathrm{N}} \mathcal{N}\left(\mathrm{Y}_n \middle| \mathrm{W}\mathrm{X}_n, \sigma_0^2 \mathbb{I}\right) d\mathrm{W} \\ &= \int \prod_{l=1}^{\mathrm{L}} \mathcal{N}\left(\mathrm{W}_{l,:}^{\mathbb{T}} \middle| \vec{0}, \alpha^{-1}\mathbb{I}\right) \prod_{l=1}^{\mathrm{L}} \mathcal{N}\left(\mathrm{Y}_{:,l} \middle| \mathrm{X}\mathrm{W}_{l,:}^{\mathbb{T}}, \sigma_0^2 \mathbb{I}\right) d\mathrm{W} \\ &= \prod_{l=1}^{\mathrm{L}} \mathcal{N}\left(\mathrm{Y}_{:,l} \middle| \vec{0}, \sigma_0^2 \mathbb{I} + \alpha \mathrm{X}\mathrm{X}^{\mathbb{T}}\right) \end{aligned} \tag{B.76}$$

Note that $\sigma_0^2 \mathbb{I} + \alpha \mathrm{X}\mathrm{X}^{\mathbb{T}} = \mathbb{K}_{\mathcal{X}\mathcal{X}}$ for the sum of constant and linear kernels $\mathbb{K} = \mathbb{K}_{\text{const}} + \mathbb{K}_{\text{linear}}$ in (B.4) and (B.19) with $\Sigma = \alpha\mathbb{I}$ and $\varphi\left(u\right) \equiv u$: this approach is called the kernel trick [Theodor-

idis & Koutroumbas (2008)]. By perceiving (B.76) as a special case where the linear kernel is chosen, the GPLVM has the general objective function for an arbitrary kernel $\mathbb{K}$ as follows:

$$p\left(\mathcal{Y}|\mathcal{X}\right) = \prod_{l=1}^{\mathrm{L}} \mathcal{N}\left(\mathrm{Y}_{:,l}|\vec{0}, \mathbb{K}_{\mathcal{X}\mathcal{X}}\right) \propto |\mathbb{K}_{\mathcal{X}\mathcal{X}}|^{-\mathrm{L}/2} \exp\left(-\frac{1}{2}\mathrm{trace}\left(\mathbb{K}_{\mathcal{X}\mathcal{X}}^{-1}\mathcal{Y}\mathcal{Y}^{\mathbb{T}}\right)\right) \tag{B.77}$$

Now the goal is to find $\mathcal{X}$ and kernel hyperparameters of $\mathbb{K}$ that maximize (B.77) as the objective function: the maximizer $\mathcal{X}$ is then regarded as the low dimension reduction of $\mathcal{Y}$. It can be done by a gradient descent on the logarithm of (B.77) whereas (B.76) is maximized explicitly via eigenvalue methods.

B.6 Deep Gaussian Process

As discussed at the top of this section, the GP models are often suboptimal from the kernel misspecification. Because there is no gold rule for finding the optimal kernel to the given data, one cannot avoid the following dilemma: making kernels complicated has a potential to make them more misspecified whereas choosing simple kernels is itself misspecified by their simple structures.

Regression (section B.2) and classification (section B.3) models adopt a column vector (let us call it as a one-dimensional hidden layer) as β while state-space models (section B.4) allow it to have a matrix form (a multi-dimensional hidden layer). Regardless of the associations, however, each model covered so far assumes only one hidden state β in the model (call them 'shallow' DP models) If the model with one hidden layer is working, why is the model not to have more than one?

Deep Gaussian process (DGP) [Damianou & Lawrence (2013)] challenges such a drawback by stacking multiple multi-dimensional hidden layers: figure B.7 shows its example graphical representation. Each hidden layer is stacked with the very previous one as the input. For simplicity, suppose that in the DPG model there are two hidden layers, $\mathrm{N} \times \mathrm{H}_1$ and $\mathrm{N} \times \mathrm{H}_2$ matrices β and γ, with $\mathcal{X} = \{\mathrm{X}_n\}_{n=1}^{\mathrm{N}} \subseteq \mathbb{R}^{\mathrm{D}}$ and $\mathcal{Y} = \{\mathrm{Y}_n\}_{n=1}^{\mathrm{N}} \subseteq \mathbb{R}^{\mathrm{L}}$ as the input and output of the training

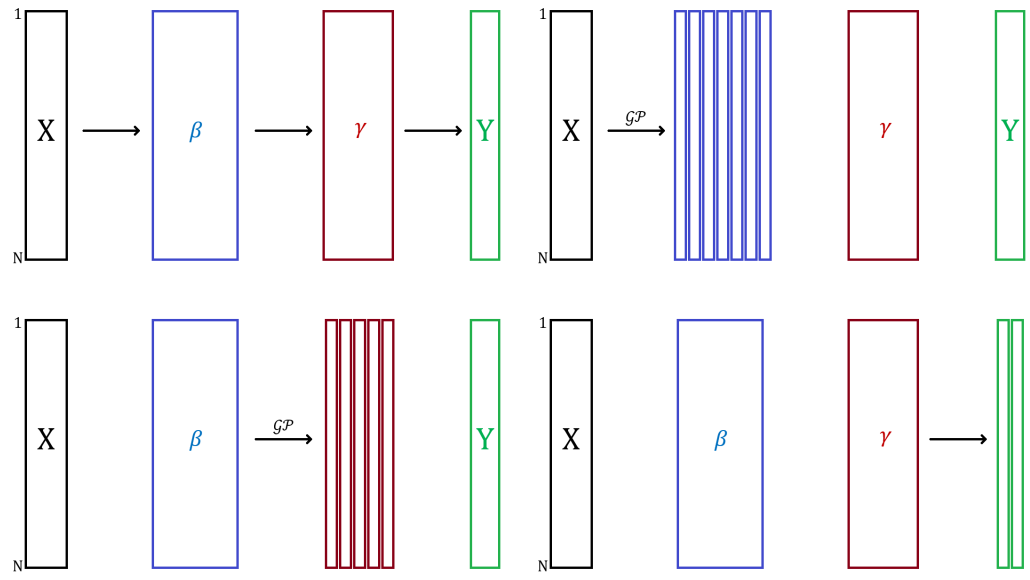


Figure B.7: A graphical representation of DGP. X and Y are the training inputs and outputs, respectively, and β and γ are the latent states (upper-left). Each column vector of β is assumed to follow a GP with X as the input (upper-right). Each column vector of γ is then assumed to follow another GP with β as the input (lower-left). Finally, Y is modelled given γ (lower-right).

data, as usual. Then, the priors on β and γ are given as follows:

$$\begin{aligned} p\left(\beta|\mathcal{X}\right) &= \prod_{h=1}^{\mathrm{H}_1} \mathcal{GP}\left(\beta_{:,h}\middle|\vec{0}, \mathbb{K}^{(1)}_{\mathcal{X}\mathcal{X}}\right) \\ p\left(\gamma|\beta\right) &= \prod_{h=1}^{\mathrm{H}_2} \mathcal{GP}\left(\gamma_{:,h}\middle|\vec{0}, \mathbb{K}^{(2)}_{\beta\beta}\right) \end{aligned} \tag{B.78}$$

With the observational model $p\left(\mathcal{Y}|\gamma\right)$, the posterior of hidden layers and posterior predictive (B.79) are given as follows:

$$\begin{aligned} &p\left(\beta,\gamma|\mathcal{X},\mathcal{Y}\right) \propto p\left(\mathcal{Y}|\gamma\right) p\left(\gamma|\beta\right) p\left(\beta|\mathcal{X}\right) \\ &p\left(y|x,\mathcal{X},\mathcal{Y}\right) = \int p\left(y|c\right) \iiint p\left(c|\gamma,b,\beta\right) p\left(b|\beta,x,\mathcal{X}\right) p\left(\beta,\gamma|\mathcal{X},\mathcal{Y}\right) d\beta d\gamma db dc \end{aligned} \tag{B.79}$$

Due to the nonlinearity of β in (B.78), the forms (B.79) cannot be expressed in closed forms. We instead consider the variational inference. Let $q\left(\beta,\gamma\right) \triangleq q_1\left(\beta\right) q_2\left(\gamma\right)$ the variation distribution to approximate $p\left(\beta,\gamma|\mathcal{X},\mathcal{Y}\right)$. Then, the following ELBO $\mathcal{L}$ is derived from the nonnegative

KL divergence $\mathbb{D}_{\mathrm{KL}}\left(q||p\left(\cdot|\mathcal{X},\mathcal{Y}\right)\right)$:

$$\log p\left(\mathcal{Y}|\mathcal{X}\right) \geq \mathbb{E}_{q_2(\gamma)}\left[\log p\left(\mathcal{Y}|\gamma\right)\right] - \mathbb{E}_{q_1(\beta)}\left[\mathbb{D}_{\mathrm{KL}}\left(q_2||p\left(\cdot|\beta\right)\right)\right] - \mathbb{D}_{\mathrm{KL}}\left(q_1||p\left(\cdot|\mathcal{X}\right)\right) = \mathcal{L} \quad \text{(B.80)}$$

Note that the KL divergences in (B.80) can be expressed in closed forms if q_1 and q_2 are given to be Gaussian. Once KL divergence are given explicitly, the reparameterization trick assists in computing the numerical gradient of $\mathcal{L}$. The variational distribution q can replace $p\left(\beta,\gamma|\mathcal{X},\mathcal{Y}\right)$ in (B.79) for the approximation after learning.

Details of the straightforward extension of the above formulations to the cases where the number of hidden layers is bigger than 2 with more complicated variational distributions can be found in [Damianou & Lawrence (2013); Salimbeni & Deisenroth (2017)]: they also deal with the cases where the input $\mathcal{X}$ is also assumed to be latent.

B.7 Variational Inference

So far, what we have covered are the exact modelling, i.e., all arithmetic operations are assumed to be feasible. In practice, however, matrix inversion quickly becomes intractable as the size of the matrix increases, as discussed earlier. Thus, the GP models have been limited and moreover often misunderstood as an anachronism in the era of big data.

Various attempts have been made to process big data with GP-based models: [Liu et al. (2020)] reviews them briefly. This subsection summarizes one of them, the variational free energy (VFE) proposed by [Titsias (2009)], as a variational approach. Note that all GP models that have been introduced in sections B.2 to B.4 can be supplemented by this VFE technique, thus modifiable to the tractable models, regardless of the data size: most of the cited publications assign parts to demonstrate such extensions. The reason why such extensions have been intentionally ignored in the previous subsections is to avoid the distraction that prevents the readers from grasping the essence of the GP models.

The core idea of VFE is assuming additional pair of pseudo-inputs $\mathcal{Z}$ and the associated latent

state γ that follows a GP with an input $\mathcal{Z}$ where γ is the sufficient statistic for β. With that pair, we have the following extended formulation of the posterior predictive:

$$\begin{aligned} p\left(y|x,\mathcal{X},\mathcal{Y}\right) &= \int p\left(y|x,b\right) \iint p\left(b|\beta,\gamma,x,\mathcal{X},\mathcal{Z}\right) p\left(\beta|\gamma,\mathcal{X},\mathcal{Z}\right) p\left(\gamma|\mathcal{X},\mathcal{Z},\mathcal{Y}\right) d\gamma d\beta db \\ &= \int p\left(y|x,b\right) \iint p\left(b|\gamma,x,\mathcal{Z}\right) p\left(\beta|\gamma,\mathcal{X},\mathcal{Z}\right) p\left(\gamma|\mathcal{X},\mathcal{Z},\mathcal{Y}\right) d\gamma d\beta db \\ &= \int p\left(y|x,b\right) \int p\left(b|\gamma,x,\mathcal{Z}\right) p\left(\gamma|\mathcal{X},\mathcal{Z},\mathcal{Y}\right) d\gamma db \end{aligned} \tag{B.81}$$

Now the problem is converted to computing $p\left(\gamma|\mathcal{X},\mathcal{Z},\mathcal{Y}\right)$ that is also an uneasy work. One resolution is to utilize the variational distribution $q\left(\beta,\gamma|\mathcal{X},\mathcal{Z}\right) = p\left(\beta|\gamma,\mathcal{X},\mathcal{Z}\right)\varphi\left(\gamma\right)$ that approximates $p\left(\beta,\gamma|\mathcal{X},\mathcal{Z},\mathcal{Y}\right)$. The following ELBO is then derived from the nonnegative KL divergence $\mathbb{D}_{\mathrm{KL}}\left(q\left(\cdot|\mathcal{X},\mathcal{Z}\right)||p\left(\cdot|\mathcal{X},\mathcal{Z},\mathcal{Y}\right)\right)$:

$$\begin{aligned} \log p\left(\mathcal{Y}|\mathcal{X}\right) &\geq \int \varphi\left(\gamma\right) \left(\int p\left(\beta|\gamma,\mathcal{X},\mathcal{Z}\right) \log p\left(\mathcal{Y}|\mathcal{X},\beta\right) d\beta - \log \frac{\varphi\left(\gamma\right)}{p\left(\gamma|\mathcal{Z}\right)} \right) d\gamma \\ &= \int \varphi\left(\gamma\right) \int p\left(\beta|\gamma,\mathcal{X},\mathcal{Z}\right) \log p\left(\mathcal{Y}|\mathcal{X},\beta\right) d\beta d\gamma - \mathbb{D}_{\mathrm{KL}}\left(\varphi||p\left(\cdot|\mathcal{Z}\right)\right) \end{aligned} \tag{B.82}$$

In the regression setting that assumes $p\left(\mathcal{Y}|\mathcal{X},\beta\right)$ to follow (B.25), [Titsias (2009)] shows that the optimal φ that maximizes (B.82) can be derived explicitly and is given as follows:

$$\begin{aligned} \varphi\left(\gamma\right) &= \mathcal{N}\left(\gamma|\ddot{\mu},\ddot{\Sigma}\right) \\ \ddot{\mu} &= \mu_{\mathcal{Z}} + \mathbb{K}_{\mathcal{Z}\mathcal{X}}\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{Z}}\mathbb{K}_{\mathcal{Z}\mathcal{Z}}^{-1}\mathbb{K}_{\mathcal{Z}\mathcal{X}}\right)^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}\right) \\ \ddot{\Sigma} &= \mathbb{K}_{\mathcal{Z}\mathcal{Z}} - \mathbb{K}_{\mathcal{Z}\mathcal{X}}\left(\Lambda_{\mathcal{X}} + \mathbb{K}_{\mathcal{X}\mathcal{Z}}\mathbb{K}_{\mathcal{Z}\mathcal{Z}}^{-1}\mathbb{K}_{\mathcal{Z}\mathcal{X}}\right)^{-1}\mathbb{K}_{\mathcal{X}\mathcal{Z}} \end{aligned} \tag{B.83}$$

Because $q\left(\gamma|\mathcal{X},\mathcal{Z}\right) = \int q\left(\beta,\gamma|\mathcal{X},\mathcal{Z}\right) d\beta = \varphi\left(\gamma\right)$ approximates $p\left(\gamma|\mathcal{X},\mathcal{Z},\mathcal{Y}\right)$, the posterior predictive (B.81) is given by the following closed form:

$$\begin{aligned} p\left(y|x,\mathcal{X},\mathcal{Y}\right) &= \mathcal{N}\left(y|m\left(x\right), \Lambda_x + s\left(x\right)\right) \\ m\left(x\right) &= \mu_x + \mathbb{K}_{x\mathcal{Z}}\left(\mathbb{K}_{\mathcal{Z}\mathcal{Z}} + \mathbb{K}_{\mathcal{Z}\mathcal{X}}\Lambda_{\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}\mathcal{Z}}\right)^{-1}\mathbb{K}_{\mathcal{Z}\mathcal{X}}\Lambda_{\mathcal{X}}^{-1}\left(\mathcal{Y} - \mu_{\mathcal{X}}\right) \\ s\left(x\right) &= \mathbb{K}_{xx} - \mathbb{K}_{x\mathcal{Z}}\mathbb{K}_{\mathcal{Z}\mathcal{Z}}^{-1}\mathbb{K}_{\mathcal{Z}x} + \mathbb{K}_{x\mathcal{Z}}\left(\mathbb{K}_{\mathcal{Z}\mathcal{Z}} + \mathbb{K}_{\mathcal{Z}\mathcal{X}}\Lambda_{\mathcal{X}}^{-1}\mathbb{K}_{\mathcal{X}\mathcal{Z}}\right)^{-1}\mathbb{K}_{\mathcal{Z}x} \end{aligned} \tag{B.84}$$

Note that $\Lambda_{\mathcal{X}}^{-1}$ is efficiently computed regardless of $|\mathcal{X}|$ since here $\Lambda_{\mathcal{X}}$ is a diagonal matrix.

If $p(\mathcal{Y}|\mathcal{X},\beta)$ is not Gaussian, then φ is not optimized in a closed form. Thus, one needs to consider the reparameterization trick to efficiently approximate the gradient of ELBO in (B.82), after parameterizing φ.

Though [Bauer et al. (2016)] process the comparison and recommend the VFE rather than the fully independent training conditional (FITC) [Csató & Opper (2002); Quiñonero-Candela & Rasmussen (2005)] for dealing with huge datasets, VFE has a certain drawback: the performance is depending on the choice of the pseudo-inputs $\mathcal{Z}$. With an ill-posed $\mathcal{Z}$, ELBO in (B.82) could be deviated far from the exact log-ML $\log p(\mathcal{Y}|\mathcal{X})$. In principle, optimizing $\mathcal{Z}$ is not straightforward. [Bui et al. (2017)] attempt both of the pseudo-input optimization and the unification of VFE and FITC by power expectation propagation.

References

Abramowitz, M. & Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.

Ades, M. & van Leeuwen, P. J. (2013). An exploration of the equivalent weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, 139(672), 820–840.

Adkins, J. F., McIntyre, K., & Schrag, D. P. (2002). The salinity, temperature, and δ^{18}o of the glacial deep ocean. *Science*, 298(5599), 1769–1773.

Ahn, S. (2016). *Bayesian Inference in Statistical Analysis of Paleoclimate Records*. PhD thesis, Brown University.

Arz, H. W., Pätzold, J., Müller, P. J., & Moammar, M. O. (2003). Influence of northern hemisphere climate and global sea level rise on the restricted red sea marine environment during termination i. *Paleoceanography*, 18(2).

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Bard, E. (2001). Comparison of alkenone estimates with other paleotemperature proxies. *Geochemistry, Geophysics, Geosystems*, 2(1).

Bard, E., Rostek, F., & Ménot-Combes, G. (2004). Radiocarbon calibration beyond 20,000 ^{14}c yr b.p. by means of planktonic foraminifera of the iberian margin. *Quaternary Research*, 61(2), 204–214.

Bard, E., Rostek, F., Turon, J.-L., & Gendreau, S. (2000). Hydrological impact of heinrich events in the subtropical northeast atlantic. *Science*, 289(5483), 1321–1324.

Barron, J. A., Heusser, L., Herbert, T., & Lyle, M. (2003). High-resolution climatic evolution of coastal northern california during the past 16,000 years. *Paleoceanography*, 18(1).

Barrows, T. T., Juggins, S., De Deckker, P., Calvo, E., & Pelejero, C. (2007). Long-term sea surface temperature and climate change in the australian–new zealand region. *Paleoceanography*, 22(2).

Bauer, M., van der Wilk, M., & Rasmussen, C. E. (2016). Understanding probabilistic sparse gaussian process approximations. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16 (pp. 1533–1541). USA: Curran Associates Inc.

Becker, K. W., Lipp, J. S., Versteegh, G. J., Wörmer, L., & Hinrichs, K.-U. (2015). Rapid and simultaneous analysis of three molecular sea surface temperature proxies and application to sediments from the sea of marmara. *Organic Geochemistry*, 85, 42 – 53.

Ben Abdessalem, A., Dervilis, N., Wagg, D., & Worden, K. (2019). Model selection and parameter estimation of dynamical systems using a novel variant of approximate bayesian computation. *Mechanical Systems and Signal Processing*, 122, 364 – 386.

Bendle, J. & Rosell-Mele, A. (2004). Distributions of uk37 and uk37' in the surface waters and sediments of the nordic seas: Implications for paleoceanography. *Geochemistry Geophysics Geosystems*.

Bendle, J. A. P. & Rosell-Melé, A. (2007). High-resolution alkenone sea surface temperature variability on the north icelandic shelf: implications for nordic seas palaeoclimatic development during the holocene. *The Holocene*, 17(1), 9–24.

Benthien, A. & Müller, P. J. (2000). Anomalously low alkenone temperatures caused by lateral particle and sediment transport in the malvinas current region, western argentine basin. *Deep Sea Research Part I: Oceanographic Research Papers*, 47(12), 2369 – 2393.

Bierens, H. J. (1994). *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge University Press.

Binois, M., Gramacy, R. B., & Ludkovski, M. (2018). Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4), 808–821.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Blaauw, M. & Christen, J. A. (2005). Radiocarbon peat chronologies and environmental change. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54(4), 805–816.

Blaauw, M. & Christen, J. A. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Anal.*, 6(3), 457–474.

Blomqvist, K., Kaski, S., & Heinonen, M. (2019). Deep convolutional gaussian process. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Bodin, E., Campbell, N. D. F., & Ek, C. H. (2017). Latent gaussian process regression. arXiv:1707.05534v2 [stat.ML].

Bowen, G. J., Fischer-Femal, B., Reichart, G.-J., Sluijs, A., & Lear, C. H. (2020). Joint inversion of proxy system models to reconstruct paleoenvironmental time series from heterogeneous data. *Climate of the Past*, 16(1), 65–78.

Briers, M., Doucet, A., & Maskell, S. (2010). Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62(1), 61–89.

Bui, T. D., Yan, J., & Turner, R. E. (2017). A unifying framework for gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18(104), 1–72.

Cacho, I., Grimalt, J. O., Canals, M., Sbaffi, L., Shackleton, N. J., Schönfeld, J., & Zahn, R. (2001). Variability of the western mediterranean sea surface temperature during the last 25,000 years and its connection with the northern hemisphere climatic changes. *Paleoceanography*, 16(1), 40–52.

Cacho, I., Grimalt, J. O., Pelejero, C., Canals, M., Sierro, F. J., Flores, J. A., & Shackleton, N. (1999). Dansgaard-oeschger and heinrich event imprints in alboran sea paleotemperatures. *Paleoceanography*, 14(6), 698–705.

Calandra, R., Peters, J., Rasmussen, C. E., & Deisenroth, M. P. (2016). Manifold gaussian processes for regression. In *International Joint Conference on Neural Networks (IJCNN)* (pp. 3338–3345).: IEEE.

Calvo, E., Pelejero, C., De Deckker, P., & Logan, G. A. (2007). Antarctic deglacial pattern in a 30 kyr record of sea surface temperature offshore south australia. *Geophysical Research Letters*, 34(13).

Calvo, E., Pelejero, C., Herguera, J. C., Palanques, A., & Grimalt, J. O. (2001). Insolation dependence of the southeastern subtropical pacific sea surface temperature over the last 400 kyrs. *Geophysical Research Letters*, 28(12), 2481–2484.

Caniupán, M., Lamy, F., Lange, C. B., Kaiser, J., Arz, H., Kilian, R., Baeza Urrea, O., Aracena, C., Hebbeln, D., Kissel, C., Laj, C., Mollenhauer, G., & Tiedemann, R. (2011). Millennial-scale sea surface temperature and patagonian ice sheet changes off southernmost chile (53°s) over the past ~60 kyr. *Paleoceanography*, 26(3).

Castañeda, I. S., Schefuß, E., Pätzold, J., Sinninghe Damsté, J. S., Weldeab, S., & Schouten, S. (2010). Millennial-scale sea surface temperature changes in the eastern mediterranean (nile river delta region) over the last 27,000 years. *Paleoceanography*, 25(1).

Catanzaro, E. J., Champion, C. E., Garner, E. L., Marinenko, G., Sappenfield, K. M., & Shields, W. R. (1970). *Standard Reference Materials: Boric acid; isotopic, and assay standard reference materials*. National Bureau of Standards (U.S.).

Chapman, M. R., Shackleton, N. J., Zhao, M., & Eglinton, G. (1996). Faunal and alkenone reconstructions of subtropical north atlantic surface hydrography and paleotemperature over the last 28 kyr. *Paleoceanography*, 11(3), 343–357.

Chazen, C. (2011). *Holocene climate evolution of the eastern tropical Pacific told from high resolution climate records from the Peru margin and equatorial upwelling regions*. PhD thesis, Brown University.

Chen, R. & Liu, J. S. (2000). Mixture kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3), 493–508.

Chen, W., Mohtadi, M., Schefuß, E., & Mollenhauer, G. (2014). Organic-geochemical proxies of sea surface temperature in surface sediments of the tropical eastern indian ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, 88, 17 – 29.

Cho, Y. & Saul, L. K. (2009). Kernel methods for deep learning. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09 (pp. 342–350). Red Hook, NY, USA: Curran Associates Inc.

Christen, J. & Sergio, E. (2009). A new robust statistical model for radiocarbon data. *Radiocarbon*, 51(3), 1047–1059.

Clifford, P. (1990). Markov random fields in statistics. In G. Grimmett & D. Welsh (Eds.), *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley* (pp. 19–32). Oxford: Oxford University Press.

Collins, J., Schefuß, E., Heslop, D., Mulitza, S., Prange, M., Zabel, M., Tjallingii, R., Dokken, T., Huang, E., Mackensen, A., Schulz, M., Tian, J., Zarriess, M., & Wefer, G. (2011). Interhemispheric symmetry of the tropical african rainbelt over the past 23,000 years. *Nature Geoscience*, 4, 42–45.

Conte, M. H., Sicre, M.-A., Rühlemann, C., Weber, J. C., Schulte, S., Schulz-Bull, D., & Blanz, T. (2006). Global temperature calibration of the alkenone unsaturation index (uk′37) in surface waters and comparison with surface sediments. *Geochemistry, Geophysics, Geosystems*, 7(2).

Csató, L. & Opper, M. (2002). Sparse on-line gaussian processes. *Neural Computation*, 14(3), 641–668.

Damianou, A. & Lawrence, N. (2013). Deep gaussian processes. In C. M. Carvalho & P. Ravikumar (Eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research* (pp. 207–215). Scottsdale, Arizona, USA: PMLR.

Damianou, A. C., Titsias, M. K., & Lawrence, N. D. (2016). Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research*, 17(42), 1–62.

Demenocal, P., Ortiz, J., Guilderson, T., Adkins, J., Sarnthein, M., Baker, L., & Yarusinsky, M. (2000). Abrupt onset and termination of the african humid period. *Quaternary Science Reviews - QUATERNARY SCI REV*, 19, 347–361.

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

Dickson, A. (2010). The carbon dioxide system in seawater: Equilibrium chemistry and measurements. *Guide to Best Practices for Ocean Acidification Research and Data Reporting*, (pp. 17–40).

Diz, P., Hall, I. R., Zahn, R., & Molyneux, E. G. (2007). Paleoceanography of the southern agulhas plateau during the last 150 ka: Inferences from benthic foraminiferal assemblages and multispecies epifaunal carbon isotopes. *Paleoceanography*, 22(4).

Doose, H., Prahl, F. G., & Lyle, M. W. (1997). Biomarker temperature estimates for modern and last glacial surface waters of the california current system between 33° and 42°n. *Paleoceanography*, 12(4), 615–622.

Doucet, A., de Freitas, N., & Gordon, N. E. (2001). *Sequential Monte Carlo methods in practice*. Springer.

Duane, S., Kennedy, A., Pendleton, B., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195, 216–222.

Dubois, N., Kienast, M., Normandeau, C., & Herbert, T. D. (2009). Eastern equatorial pacific cold tongue during the last glacial maximum as seen from alkenone paleothermometry. *Paleoceanography*, 24(4).

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13 (pp. III–1166–III–1174).: JMLR.org.

Duvenaud, D. K. (2014). *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge.

Dyez, K. A., Hönisch, B., & Schmidt, G. A. (2018). Early pleistocene obliquity-scale pco2 variability at 1.5 million years ago. *Paleoceanography and Paleoclimatology*, 33(11), 1270–1291.

Eggleton, T. (2012). *A Short Introduction to Climate Change*. Cambridge University Press.

Eleftheriadis, S., Nicholson, T. F., Deisenroth, M. P., & Hensman, J. (2017). Identification of gaussian process state space models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (pp. 5315–5325). USA: Curran Associates Inc.

Emeis, K.-C., Struck, U., Schulz, H.-M., Rosenberg, R., Bernasconi, S., Erlenkeuser, H., Sakamoto, T., & Martinez-Ruiz, F. (2000). Temperature and salinity variations of mediterranean sea surface waters over the last 16,000 years from records of planktonic stable oxygen isotopes and alkenone unsaturation ratios. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 158(3), 259 – 280.

Epstein, S., Buchsbaum, R., Lowenstam, H. A., & Urey, H. C. (1953). Revised carbonate-water isotopic temperature scale. *GSA Bulletin*, 64(11), 1315–1326.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5), 10143–10162.

Fallet, U., Castañeda, I. S., Henry-Edwards, A., Richter, T. O., Boer, W., Schouten, S., & Brummer, G.-J. (2012). Sedimentation and burial of organic and inorganic temperature proxies in the mozambique channel, sw indian ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, 59, 37 – 53.

Filippova, A., Kienast, M., Frank, M., & Schneider, R. R. (2016). Alkenone paleothermometry in the north atlantic: A review and synthesis of surface sediment data and calibrations. *Geochemistry, Geophysics, Geosystems*, 17(4), 1370–1382.

Foster, G. L. & Rae, J. W. (2016). Reconstructing ocean ph with boron isotopes in foraminifera. *Annual Review of Earth and Planetary Sciences*, 44(1), 207–237.

Fraser, N., Kuhnt, W., Holbourn, A., Bolliet, T., Andersen, N., Blanz, T., & Beaufort, L. (2014). Precipitation variability within the west pacific warm pool over the past 120 ka: Evidence from the davao gulf, southern philippines. *Paleoceanography*, 29(11), 1094–1110.

Frigola, R., Chen, Y., & Rasmussen, C. E. (2014). Variational gaussian process state-space models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14 (pp. 3680–3688). Cambridge, MA, USA: MIT Press.

Garcia-Olivares, A. & Herrero, C. (2012). Fitting the last pleistocene d 18 o and co 2 time series with simple box models. *Scientia Marina*, 76S1, 209–218.

Garcia-Olivares, A. & Herrero, C. (2013). Simulation of glacial-interglacial cycles by simple relaxation models: Consistency with observational results. *Climate Dynamics*, 41, 1307–1331.

Gebbie, G. & Huybers, P. (2012). The mean age of ocean waters inferred from radiocarbon observations: Sensitivity to surface sources and accounting for mixing histories. *Journal of Physical Oceanography*, 42(2), 291–305.

Geenens, G. (2011). Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys*, 5, 30–43.

Gelman, A. (2008). Objections to bayesian statistics. *Bayesian Analysis*, 3(3), 445–449.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741.

Genton, M. G. (2002). Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2, 299–312.

Gibbs, M. N. (1997). *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge.

Girolami, M. & Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2), 123–214.

Goldberg, P. W., Williams, C. K. I., & Bishop, C. M. (1998). Regression with input-dependent noise: A gaussian process treatment. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*, NIPS '97 (pp. 493–499). Cambridge, MA, USA: MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Grousset, F. (1993). Patterns of ice-rafted detritus in the glacial north atlantic (40- 55°n). *Paleoceanography*, 8, 175–192.

Gutiérrez, D., Bouloubassi, I., Sifeddine, A., Purca, S., Goubanova, K., Graco, M., Field, D., Méjanelle, L., Velazco, F., Lorre, A., Salvatteci, R., Quispe, D., Vargas, G., Dewitte, B., & Ortlieb, L. (2011). Coastal cooling and increased productivity in the main upwelling zone off peru since the mid-twentieth century. *Geophysical Research Letters*, 38(7).

Haji Ghassemi, N. (2013). Analytic long-term forecasting with periodic gaussian processes. Master's thesis, Blekinge Institute of Technology.

Hangos, K., Szederkényi, G., Lakner, R., & Gerzson, M. (2001). *Intelligent Control Systems: An Introduction with Examples*. Applied Optimization. Springer.

Harada, N., Ahagon, N., Sakamoto, T., Uchida, M., Ikehara, M., & Shibata, Y. (2006). Rapid fluctuation of alkenone temperature in the southwestern okhotsk sea during the past 120 ky. *Global and Planetary Change*, 53(1), 29 – 46. Late Quaternary paleoceanography of the northwestern Pacific: Results from IMAGES program.

Haslett, J. & Parnell, A. (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(4), 399–418.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

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

Hemming, N. & Hanson, G. (1992). Boron isotopic composition and concentration in modern marine carbonates. *Geochimica et Cosmochimica Acta*, 56(1), 537 – 543.

Henehan, M. J., Foster, G. L., Bostock, H. C., Greenop, R., Marshall, B. J., & Wilson, P. A. (2016). A new boron isotope-ph calibration for orbulina universa, with implications for understanding and accounting for 'vital effects'. *Earth and Planetary Science Letters*, 454, 282 – 292.

Herbert, T. D., Schuffert, J. D., Andreasen, D., Heusser, L., Lyle, M., Mix, A., Ravelo, A. C., Stott, L. D., & Herguera, J. C. (2001). Collapse of the california current during glacial maxima linked to climate change on land. *Science*, 293(5527), 71–76.

Herbert, T. D., Schuffert, J. D., Thomas, D., Lange, C., Weinheimer, A., Peleo-Alampay, A., & Herguera, J.-C. (1998). Depth and seasonality of alkenone production along the california margin inferred from a core top transect. *Paleoceanography*, 13(3), 263–271.

Hernández-Lobato, D., Hernández-Lobato, J. M., & Dupont, P. (2011). Robust multi-class gaussian process classification. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11 (pp. 280–288). Red Hook, NY, USA: Curran Associates Inc.

Hernández-Sánchez, M., Woodward, E., Taylor, K., Henderson, G., & Pancost, R. (2014). Variations in gdgt distributions through the water column in the south east atlantic ocean. *Geochimica et Cosmochimica Acta*, 132, 337 – 348.

Ho, S. L., Mollenhauer, G., Fietz, S., Martínez-Garcia, A., Lamy, F., Rueda, G., Schipper, K., Méheust, M., Rosell-Melé, A., Stein, R., & Tiedemann, R. (2014). Appraisal of tex86 and tex86l thermometries in subpolar and polar regions. *Geochimica et Cosmochimica Acta*, 131, 213 – 226.

Ho, S. L., Mollenhauer, G., Lamy, F., Martínez-Garcia, A., Mohtadi, M., Gersonde, R., Hebbeln, D., Nunez-Ricardo, S., Rosell-Melé, A., & Tiedemann, R. (2012). Sea surface temperature variability in the pacific sector of the southern ocean over the past 700 kyr. *Paleoceanography*, 27(4).

Ho, S. L., Yamamoto, M., Mollenhauer, G., & Minagawa, M. (2011). Core top tex86 values in the south and equatorial pacific. *Organic Geochemistry*, 42(1), 94 – 99.

Hoffman, M. D. & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47), 1593–1623.

Hogg, A. G., Hua, Q., Blackwell, P. G., Niu, M., Buck, C. E., Guilderson, T. P., Heaton, T. J., Palmer, J. G., Reimer, P. J., & Reimer, R. W. (2013). Shcal13 southern hemisphere calibration, 0–50,000 years cal bp. *Radiocarbon*, 55(4), 1889–1903.

Hönisch, B., Hemming, N. G., Archer, D., Siddall, M., & McManus, J. F. (2009). Atmospheric carbon dioxide concentration across the mid-pleistocene transition. *Science*, 324(5934), 1551–1554.

Horikawa, K., Minagawa, M., Murayama, M., Kato, Y., & Asahi, H. (2006). Spatial and temporal sea-surface temperatures in the eastern equatorial pacific over the past 150 kyr. *Geophysical Research Letters*, 33(13).

Houtekamer, P. L. & Mitchell, H. L. (1998). Data assimilation using an ensemble kalman filter technique. *Monthly Weather Review*, 126(3), 796–811.

Hu, J., Meyers, P. A., Chen, G., Peng, P., & Yang, Q. (2012). Archaeal and bacterial glycerol dialkyl glycerol tetraethers in sediments from the eastern lau spreading center, south pacific ocean. *Organic Geochemistry*, 43, 162 – 167.

Huang, C.-Y., Wu, S.-F., Zhao, M., Chen, M.-T., Wang, C.-H., Tu, X., & Yuan, P. B. (1997). Surface ocean and monsoon climate variability in the south china sea since the last glaciation. *Marine Micropaleontology*, 32(1), 71 – 94. Marine Micropaleontology in China.

Ikehara, M., Kawamura, K., Ohkouchi, N., Kimoto, K., Murayama, M., Nakamura, T., Oba, T., & Taira, A. (1997). Alkenone sea surface temperature in the southern ocean for the last two deglaciations. *Geophysical Research Letters*, 24(6), 679–682.

Indermühle, A., Monnin, E., Stauffer, B., Stocker, T. F., & Wahlen, M. (2000). Atmospheric co2 concentration from 60 to 20 kyr bp from the taylor dome ice core, antarctica. *Geophysical Research Letters*, 27(5), 735–738.

Jaakkola, T., Diekhans, M., & Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2), 95–114.

Jaeschke, A., Rühlemann, C., Arz, H., Heil, G., & Lohmann, G. (2007). Coupling of millennial-scale changes in sea surface temperature and precipitation off northeastern brazil with high-latitude climate shifts during the last glacial period. *Paleoceanography*, 22(4).

Jaeschke, A., Wengler, M., Hefter, J., Ronge, T. A., Geibert, W., Mollenhauer, G., Gersonde, R., & Lamy, F. (2017). A biomarker perspective on dust, productivity, and sea surface temperature in the pacific sector of the southern ocean. *Geochimica et Cosmochimica Acta*, 204, 120 – 139.

Jia, G., Zhang, J., Chen, J., Peng, P., & Zhang, C. L. (2012). Archaeal tetraether lipids record subsurface water temperature in the south china sea. *Organic Geochemistry*, 50, 68 – 77.

Julier, S. J. & Uhlmann, J. K. (1997). New extension of the Kalman filter to nonlinear systems. In I. Kadar (Ed.), *Signal Processing, Sensor Fusion, and Target Recognition VI*, volume 3068 (pp. 182 – 193).: International Society for Optics and Photonics SPIE.

Kaiser, J., Lamy, F., & Hebbeln, D. (2005). A 70-kyr sea surface temperature record off southern chile (ocean drilling program site 1233). *Paleoceanography*, 20(4).

Kaiser, J., Ruggieri, N., Hefter, J., Siegel, H., Mollenhauer, G., Arz, H. W., & Lamy, F. (2014). Lipid biomarkers in surface sediments from the gulf of genoa, ligurian sea (nw mediterranean sea) and their potential for the reconstruction of palaeo-environments. *Deep Sea Research Part I: Oceanographic Research Papers*, 89, 68 – 83.

Kaiser, J., Schefuß, E., Lamy, F., Mohtadi, M., & Hebbeln, D. (2008). Glacial to holocene changes in sea surface temperature and coastal vegetation in north central chile: high versus low latitude forcing. *Quaternary Science Reviews*, 27(21), 2064 – 2075.

Kaiser, M., Otte, C., Runkler, T., & Ek, C. H. (2018). Data association with gaussian processes. arXiv:1810.07158v3 [stat.ML].

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

Kennedy, J. A. & Brassell, S. C. (1992). Molecular stratigraphy of the santa barbara basin: comparison with historical records of annual climate change. *Organic Geochemistry*, 19(1), 235 – 244. Proceedings of the 15th International Meeting on Organic Geochemistry.

Kersting, K., Plagemann, C., Pfaff, P., & Burgard, W. (2007). Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07 (pp. 393–400). New York, NY, USA: ACM.

Kienast, M., Kienast, S., Calvert, S., Eglinton, T., Mollenhauer, G., François, R., & Mix, A. (2006). Eastern pacific cooling and atlantic overturning circulation during the last deglaciation. *Nature*, 443, 846–9.

Kienast, M., MacIntyre, G., Dubois, N., Higginson, S., Normandeau, C., Chazen, C., & Herbert, T. D. (2012). Alkenone unsaturation in surface sediments from the eastern equatorial pacific: Implications for sst reconstructions. *Paleoceanography*, 27(1).

Kienast, S. S. & McKay, J. L. (2001). Sea surface temperatures in the subarctic northeast pacific reflect millennial-scale climate oscillations during the last 16 kyrs. *Geophysical Research Letters*, 28(8), 1563–1566.

Kim, J.-H., Meggers, H., Rimbu, N., Lohmann, G., Freudenthal, T., Müller, P. J., & Schneider, R. R. (2007). Impacts of the North Atlantic gyre circulation on Holocene climate off northwest Africa. *Geology*, 35(5), 387–390.

Kim, J.-H., Rimbu, N., Lorenz, S. J., Lohmann, G., Nam, S.-I., Schouten, S., Rühlemann, C., & Schneider, R. R. (2004). North pacific and north atlantic sea-surface temperature variability during the holocene. *Quaternary Science Reviews*, 23(20), 2141 – 2154. Holocene climate variability - a marine perspective.

Kim, J.-H., Schneider, R. R., Hebbeln, D., Müller, P. J., & Wefer, G. (2002a). Last deglacial sea-surface temperature evolution in the southeast pacific compared to climate changes on the south american continent. *Quaternary Science Reviews*, 21(18), 2085 – 2097.

Kim, J.-H., Schneider, R. R., Mulitza, S., & Müller, P. J. (2003). Reconstruction of se trade-wind intensity based on sea-surface temperature gradients in the southeast atlantic over the last 25 kyr. *Geophysical Research Letters*, 30(22).

Kim, J.-H., Schneider, R. R., Müller, P. J., & Wefer, G. (2002b). Interhemispheric comparison of deglacial sea-surface temperature patterns in atlantic eastern boundary currents. *Earth and Planetary Science Letters*, 194(3), 383 – 393.

Kim, J.-H., Schouten, S., Hopmans, E. C., Donner, B., & Damsté], J. S. S. (2008). Global sediment core-top calibration of the tex86 paleothermometer in the ocean. *Geochimica et Cosmochimica Acta*, 72(4), 1154 – 1173.

Kim, J.-H., van der Meer, J., Schouten, S., Helmke, P., Willmott Puig, V., Sangiorgi, F., Koc, N., Hopmans, E., & Sinninghe-Damste, J. (2010). New indices and calibrations derived from the distribution of crenarchaeal isoprenoid tetraether lipids: Implications for past sea surface temperature reconstructions. *Geochimica et Cosmochimica Acta*, 74, 4639–4654.

Kim, R. A., Lee, K. E., & Bae, S. W. (2015). Sea surface temperature proxies (alkenones, foraminiferal mg/ca, and planktonic foraminiferal assemblage) and their implications in the okinawa trough. *Prog. in Earth and Planet. Sci.*, 2(43).

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. arXiv:1312.6114v10 [stat.ML].

Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics*, 46(4), 605–623.

Klaas, M., Briers, M., de Freitas, N., Doucet, A., Maskell, S., & Lang, D. (2006). Fast particle smoothing: if i had a million particles. In *ICML*.

Klochko, K., Kaufman, A. J., Yao, W., Byrne, R. H., & Tossell, J. A. (2006). Experimental measurement of boron isotope fractionation in seawater. *Earth and Planetary Science Letters*, 248(1), 276 – 285.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press.

Koutavas, A. & Sachs, J. P. (2008). Northern timing of deglaciation in the eastern equatorial pacific from alkenone paleothermometry. *Paleoceanography*, 23(4).

Kudrass, H., Hofmann, A., Doose, H., Emeis, K., & Erlenkeuser, H. (2001). Modulation and amplification of climatic changes in the Northern Hemisphere by the Indian summer monsoon during the past 80 k.y. *Geology*, 29(1), 63–66.

Lamy, F., Rühlemann, C., Hebbeln, D., & Wefer, G. (2002). High- and low-latitude climate control on the position of the southern peru-chile current during the holocene. *Paleoceanography*, 17(2), 16–1–16–10.

Landau, L. D. & Lifshitz, E. M. (2013). *Statistical Physics: Volume 5*. Elsevier Science.

Langrené, N. & Warin, X. (2019). Fast and stable multivariate kernel density estimation by fast sum updating. *Journal of Computational and Graphical Statistics*, 28(3), 596–608.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6, 1783–1816.

Lázaro-Gredilla, M. & Titsias, M. K. (2011). Variational heteroscedastic gaussian process regression. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (pp. 841–848).

Lázaro-Gredilla, M., Van Vaerenbergh, S., & Lawrence, N. D. (2012). Overlapping mixtures of gaussian processes for the data association problem. *Pattern Recognition*, 45(4), 1386–1395.

Le, Q. V., Smola, A. J., & Canu, S. (2005). Heteroscedastic gaussian process regression. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05 (pp. 489–496).: ACM.

Le Gall, F. (2014). Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, ISSAC '14 (pp. 296–303). New York, NY, USA: ACM.

Leduc, G., Schneider, R., Kim, J.-H., & Lohmann, G. (2010). Holocene and eemian sea surface temperature trends as revealed by alkenone and mg/ca paleothermometry. *Quaternary Science Reviews*, 29(7), 989 – 1004.

Leduc, G., Vidal, L., Tachikawa, K., Rostek, F., Sonzogni, C., Beaufort, L., & Bard, E. (2007). Moisture transport across central america as a positive feedback on abrupt climatic changes. *Nature*, 445, 908–11.

Lee, K. E., Bahk, J. J., & Choi, J. (2008a). Alkenone temperature estimates for the east sea during the last 190,000 years. *Organic Geochemistry*, 39(6), 741 – 753.

Lee, K. E., Kim, J.-H., Wilke, I., Helmke, P., & Schouten, S. (2008b). A study of the alkenone, tex86, and planktonic foraminifera in the benguela upwelling system: Implications for past sea surface temperature estimates. *Geochemistry, Geophysics, Geosystems*, 9(10).

Lee, K. E., Slowey, N. C., & Herbert, T. D. (2001). Glacial sea surface temperatures in the subtropical north pacific: A comparison of u37 k′, δ18o, and foraminiferal assemblage temperature estimates. *Paleoceanography*, 16(3), 268–279.

Lee, T. & Lawrence, C. E. (2019). Heteroscedastic gaussian process regression on the alkenone over sea surface temperatures. In *Proceedings of the 9th International Workshop on Climate Informatics* (pp. 269–274).

Lee, T., Lisiecki, L. E., Rand, D., Gebbie, G., & Lawrence, C. E. (2019). A multiple continuous signal alignment algorithm with gaussian process profiles and an application to paleoceanography. arXiv:1907.08738v3 [stat.AP].

Leider, A., Hinrichs, K.-U., Mollenhauer, G., & Versteegh, G. (2010). Core-top calibration of the lipid-based u37k′ and tex86 temperature proxies on the southern italian shelf (sw adriatic sea, gulf of taranto). *Earth and Planetary Science Letters*, 300, 112–124.

Lemarchand, D., Gaillardet, J., Lewin, E., & Allègre, C. J. (2002). Boron isotope systematics in large rivers: implications for the marine boron budget and paleo-ph reconstruction over the cenozoic. *Chemical Geology*, 190(1), 123 – 140. Geochemistry of Crustal Fluids-Fluids in the Crust and Chemical Fluxes at the Earth's Surface.

Lengger, S. K., Hopmans, E. C., Sinninghe Damsté, J. S., & Schouten, S. (2014). Impact of sedimentary degradation and deep water column production on gdgt abundance and distribution in surface sediments in the arabian sea: Implications for the tex86 paleothermometer. *Geochimica et Cosmochimica Acta*, 142, 386 – 399.

Li, P. & Chen, S. (2016). A review on gaussian process latent variable models. *CAAI Transactions on Intelligence Technology*, 1(4), 366–376.

Lin, L., Khider, D., Lisiecki, L. E., & Lawrence, C. E. (2014). Probabilistic sequence alignment of stratigraphic records. *Paleoceanography*, 29(10).

Lisiecki, L. E. & Lisiecki, P. A. (2002). Application of dynamic programming to the correlation of paleoclimate records. *Paleoceanography*, 17(4), 1–1–1–12.

Lisiecki, L. E. & Raymo, M. E. (2005). A pliocene-pleistocene stack of 57 globally distributed benthic δ^{18}o records. *Paleoceanography*, 20.

Lisiecki, L. E. & Stern, J. V. (2016). Regional and global benthic δ^{18}o stacks for the last glacial cycle. *Paleoceanography*, 31(10), 1368–1394.

Liu, H., Ong, Y.-S., Shen, X., & Cai, J. (2020). When gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer.

Liu, Z. & Herbert, T. (2004). High-latitude influence on the eastern equatorial pacific climate in the early pleistocene epoch. *Nature*, 427, 720–723.

Liu, Z., Pagani, M., Zinniker, D., DeConto, R., Huber, M., Brinkhuis, H., Shah, S. R., Leckie, R. M., & Pearson, A. (2009). Global cooling during the eocene-oligocene climate transition. *Science*, 323(5918), 1187–1190.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *Journal of Machine Learning Research*, 2, 419–444.

Loftsgaarden, D. O. & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36(3), 1049–1051.

Lopes dos Santos, R. A., Prange, M., Castañeda, I. S., Schefuß, E., Mulitza, S., Schulz, M., Niedermeyer, E. M., Sinninghe Damsté, J. S., & Schouten, S. (2010). Glacial–interglacial variability in atlantic meridional overturning circulation and thermocline adjustments in the tropical north atlantic. *Earth and Planetary Science Letters*, 300(3), 407 – 414.

Lü, X., Yang, H., Song, J., Versteegh, G. J., Li, X., Yuan, H., Li, N., Yang, C., Yang, Y., Ding, W., & Xie, S. (2014). Sources and distribution of isoprenoid glycerol dialkyl glycerol tetraethers (gdgts) in sediments from the east coastal sea of china: Application of gdgt-based paleothermometry to a shallow marginal sea. *Organic Geochemistry*, 75, 24 – 35.

Lückge, A., Mohtadi, M., Rühlemann, C., Scheeder, G., Vink, A., Reinhardt, L., & Wiedicke, M. (2009). Monsoon versus ocean circulation controls on paleoenvironmental conditions off southern sumatra during the past 300,000 years. *Paleoceanography*, 24(1).

Lüthi, D., Floch, M., Bereiter, B., Blunier, T., Barnola, J.-M., Siegenthaler, U., Raynaud, D., Jouzel, J., Fischer, H., Kawamura, K., & Stocker, T. (2008). High-resolution carbon dioxide concentration record 650,000-800,000 years before present. *Nature*, 453, 379–82.

MacKay, D. J. (1998). Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168, 133–166.

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

Madureira, L. A. S., Conte, M. H., & Eglinton, G. (1995). Early diagenesis of lipid biomarker compounds in north atlantic sediments. *Paleoceanography*, 10(3), 627–642.

Marchal, O., Cacho, I., Stocker, T. F., Grimalt, J. O., Calvo, E., Martrat, B., Shackleton, N., Vautravers, M., Cortijo, E., van Kreveld, S., Andersson, C., Koç, N., Chapman, M., Sbaffi, L., Duplessy, J.-C., Sarnthein, M., Turon, J.-L., Duprat, J., & Jansen, E. (2002). Apparent long-term cooling of the sea surface in the northeast atlantic and mediterranean during the holocene. *Quaternary Science Reviews*, 21(4), 455 – 483.

Martino, L., Read, J., & Luengo, D. (2015). Independent doubly adaptive rejection metropolis sampling within gibbs sampling. *IEEE Transactions on Signal Processing*, 63(12), 3123–3138.

Martrat, B., Grimalt, J. O., Shackleton, N. J., de Abreu, L., Hutterli, M. A., & Stocker, T. F. (2007). Four climate cycles of recurring deep and surface water destabilizations on the iberian margin. *Science*, 317(5837), 502–507.

Matérn, B. (1986). *Spatial Variation*. Springer-Verlag.

Max, L., Riethdorf, J.-R., Tiedemann, R., Smirnova, M., Lembke-Jene, L., Fahl, K., Nürnberg, D., Matul, A., & Mollenhauer, G. (2012). Sea surface temperature variability and sea-ice extent in the subarctic northwest pacific during the past 15,000 years. *Paleoceanography*, 27(3).

Maybeck, P. (1979). *Stochastic Models, Estimation and Control*. Number pt. 1 in Mathematics in science and engineering. Academic Press.

McCaffrey, M. A., Farrington, J. W., & Repeta, D. J. (1990). The organic geochemistry of peru margin surface sediments: I. a comparison of the c37 alkenone and historical el niño records. *Geochimica et Cosmochimica Acta*, 54(6), 1671 – 1682.

McElhoe, B. A. (1966). An assessment of the navigation and course corrections for a manned flyby of mars or venus. *IEEE Transactions on Aerospace and Electronic Systems*, AES-2(4), 613–623.

McManus, J. F., Francois, R., Gherardi, J.-M., Keigwin, L. D., & Brown-Leger, S. (2004). Collapse and rapid resumption of atlantic meridional circulation linked to deglacial climate changes. *Nature*, 428, 834–837.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087–1092.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT.

Mohtadi, M., Oppo, D. W., Lückge, A., DePol-Holz, R., Steinke, S., Groeneveld, J., Hemme, N., & Hebbeln, D. (2011). Reconstructing the thermal structure of the upper ocean: Insights from planktic foraminifera shell chemistry and alkenones in modern sediments of the tropical eastern indian ocean. *Paleoceanography*, 26(3).

Molyneux, E. G., Hall, I. R., Zahn, R., & Diz, P. (2007). Deep water variability on the southern agulhas plateau: Interhemispheric links over the past 170 ka. *Paleoceanography*, 22(4).

Monnin, E., Indermühle, A., Dällenbach, A., Flückiger, J., Stauffer, B., Stocker, T. F., Raynaud, D., & Barnola, J.-M. (2001). Atmospheric co2 concentrations over the last glacial termination. *Science*, 291(5501), 112–114.

Mulitza, S., Prange, M., Stuut, J.-B., Zabel, M., von Dobeneck, T., Itambi, A. C., Nizou, J., Schulz, M., & Wefer, G. (2008). Sahel megadroughts triggered by glacial slowdowns of atlantic meridional overturning. *Paleoceanography*, 23(4).

Munich, M. E. & Perona, P. (1999). Continuous dynamic time warping for translation-invariant curve alignment with applications to signature verification. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1 (pp. 108–115).

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Müller, P. J., Kirst, G., Ruhland, G., [von Storch], I., & Rosell-Melé, A. (1998). Calibration of the alkenone paleotemperature index u37k′ based on core-tops from the eastern south atlantic and the global ocean (60°n-60°s). *Geochimica et Cosmochimica Acta*, 62(10), 1757 – 1772.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1), 141–142.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods*. Technical report, University of Toronto.

Neal, R. M. (2010). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54, 113–162.

Nieto-Moreno, V., Martínez-Ruiz, F., Willmott, V., García-Orellana, J., Masqué, P., & Damsté], J. S. (2013). Climate conditions in the westernmost mediterranean over the last two millennia: An integrated biomarker approach. *Organic Geochemistry*, 55, 1 – 10.

Norris, J. R. (1997). *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Nürnberg, D., Ziegler, M., Karas, C., Tiedemann, R., & Schmidt, M. W. (2008). Interacting loop current variability and mississippi river discharge over the past 400 kyr. *Earth and Planetary Science Letters*, 272(1), 278 – 289.

Ohkouchi, N., Kawamura, K., Kawahata, H., & Okada, H. (1999). Depth ranges of alkenone production in the central pacific ocean. *Global Biogeochemical Cycles - GLOBAL BIOGEOCHEM CYCLE*, 13, 695–704.

Ostertag-Henning, C. & Stax, R. (2000). Data report: Carbonate records from sites 1012, 1013, 1017, and 1019 and alkenone-based sea-surface temperatures from site 1017. *Proceedings of the Ocean Drilling Program: Scientific Results*, 167, 297–302.

Pahnke, K. & Sachs, J. P. (2006). Sea surface temperatures of southern midlatitudes 0–160 kyr b.p. *Paleoceanography*, 21(2).

Pahnke, K., Sachs, J. P., Keigwin, L., Timmermann, A., & Xie, S.-P. (2007). Eastern tropical pacific hydrologic changes during the past 27,000 years from d/h ratios in alkenones. *Paleoceanography*, 22(4).

Paillard, D. & Parrenin, F. (2004). The antarctic ice sheet and the triggering of deglaciations. *Earth and Planetary Science Letters*, 227(3), 263 – 271.

Pailler, D. & Bard, E. (2002). High frequency palaeoceanographic changes during the past 140000 yr recorded by the organic matter in sediments of the iberian margin. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 181(4), 431 – 452.

Papoulis, A. & Pillai, U. (2001). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 4th edition.

Park, Y.-H., Yamamoto, M., Nam, S.-I., Irino, T., Polyak, L., Harada, N., Nagashima, K., Khim, B.-K., Chikita, K., & Saitoh, S.-I. (2014). Distribution, source and transportation of glycerol dialkyl glycerol tetraethers in surface sediments from the western arctic ocean and the northern bering sea. *Marine Chemistry*, 165, 10 – 24.

Paterne, M., Kallel, N., Labeyrie, L., Vautravers, M., Duplessy, J.-C., Rossignol-Strick, M., Cortijo, E., Arnold, M., & Fontugne, M. (1999). Hydrological relationship between the north atlantic ocean and the mediterranean sea during the past 15-75 kyr. *Paleoceanography*, 14(5), 626–638.

Pelejero, C. & Grimalt, J. O. (1997). The correlation between the 37k index and sea surface temperatures in the warm boundary: The south china sea. *Geochimica et Cosmochimica Acta*, 61(22), 4789 – 4797.

Pelejero, C., Grimalt, J. O., Heilig, S., Kienast, M., & Wang, L. (1999). High-resolution uk 37 temperature reconstructions in the south china sea over the past 220 kyr. *Paleoceanography*, 14(2), 224–231.

Penny, S. G. & Miyoshi, T. (2016). A local particle filter for high-dimensional geophysical systems. *Nonlinear Processes in Geophysics*, 23(6), 391–405.

Peters, G. (2008). Markov chain monte carlo: stochastic simulation for bayesian inference. *Statistics in Medicine*, 27(16), 3213–3214.

Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J.-M., BASILE-DOELSCH, I., Bender, M., Chappellaz, J., Davis, M., Delaygue, G., Delmotte, M., Kotlyakov, V. M., Legrand, M., Lipenkov, V. Y., Lorius, C., Pepin, L., Ritz, C., Saltzman, E., & Stievenard, M. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*, 399(6735), 429–436.

Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., & Keogh, E. J. (2014). Dynamic time warping averaging of time series allows faster and more accurate classification. *2014 IEEE International Conference on Data Mining*, (pp. 470–479).

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3), 678–693.

Poterjoy, J. (2016). A localized particle filter for high-dimensional nonlinear systems. *Monthly Weather Review*, 144(1), 59–76.

Prahl, F., Rontani, J.-F., Zabeti, N., Walinsky, S., & Sparrow, M. (2010). Systematic pattern in u37k′ - temperature residuals for surface sediments from high latitude and other oceanographic settings. *Geochimica et Cosmochimica Acta*, 74, 131–143.

Prahl, F. G., Mix, A. C., & Sparrow, M. A. (2006). Alkenone paleothermometry: Biological lessons from marine sediment records off western south america. *Geochimica et Cosmochimica Acta*, 70(1), 101 – 117.

Prahl, F. G., Muehlhausen, L. A., & Lyle, M. (1989). An organic geochemical assessment of oceanographic conditions at manop site c over the past 26,000 years. *Paleoceanography*, 4(5), 495–510.

Prahl, F. G., Muehlhausen, L. A., & Zahnle, D. L. (1988). Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions. *Geochimica et Cosmochimica Acta*, 52(9), 2303–2310.

Quiñonero-Candela, J. & Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6, 1939–1959.

Rasmussen, C. E. & Ghahramani, Z. (2002). Infinite mixtures of gaussian process experts. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14* (pp. 881–888). MIT Press.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rauch, H. E., Tung, F., & Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8), 1445–1450.

Raynaud, D., Barnola, J.-M., Souchez, R., Lorrain, R., Petit, J.-R., Duval, P., & Lipenkov, V. Y. (2005). The record for marine isotopic stage 11. *Nature*, 436, 39–40.

Reimer, P. J., Bard, E., Bayliss, A., Beck, J. W., Blackwell, P. G., Ramsey, C. B., Buck, C. E., Cheng, H., Edwards, R. L., & Friedrich, M. (2013). Intcal13 and marine13 radiocarbon age calibration curves 0–50,000 years cal bp. *Radiocarbon*, 55(4), 1869–1887.

Rein, B., Lückge, A., Reinhardt, L., Sirocko, F., Wolf, A., & Dullo, W.-C. (2005). El niño variability off peru during the last 20,000 years. *Paleoceanography*, 20(4).

Remes, S., Heinonen, M., & Kaski, S. (2017). Non-stationary spectral kernels. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4642–4651). Curran Associates, Inc.

Richey, J. N., Hollander, D. J., Flower, B. P., & Eglinton, T. I. (2011). Merging late holocene molecular organic and foraminiferal-based geochemical records of sea surface temperature in the gulf of mexico. *Paleoceanography*, 26(1).

Rodrigo-Gámiz, M., Rampen, S. W., de Haas, H., Baas, M., Schouten, S., & Sinninghe Damsté, J. S. (2015). Constraints on the applicability of the organic temperature proxies $u^{K'}_{37}$, tex_{86} and ldi in the subpolar region around iceland. *Biogeosciences*, 12(22), 6573–6590.

Rodrigues, T., Grimalt, J. O., Abrantes, F., Naughton, F., & Flores, J.-A. (2010). The last glacial–interglacial transition (lgit) in the western mid-latitudes of the north atlantic: Abrupt sea surface temperature change and sea level implications. *Quaternary Science Reviews*, 29(15), 1853 – 1862. Special Theme: Arctic Palaeoclimate Synthesis (PP. 1674-1790).

Rodrigues, T., Grimalt, J. O., Abrantes, F. G., Flores, J. A., & Lebreiro, S. M. (2009). Holocene interdependences of changes in sea surface temperature, productivity, and fluvial inputs in the iberian continental shelf (tagus mud patch). *Geochemistry, Geophysics, Geosystems*, 10(7).

Romero, O. E., Kim, J.-H., & Donner, B. (2008). Submillennial-to-millennial variability of diatom production off mauritania, nw africa, during the last glacial cycle. *Paleoceanography*, 23(3).

Rosell-Melé, A. (1998). Interhemispheric appraisal of the value of alkenone indices as temperature and salinity proxies in high-latitude locations. *Paleoceanography*, 13, 694–703.

Rosell-Melé, A., Comes, P., Muller, P., & Ziveri, P. (2000). Alkenone fluxes and anomalous u37k′ values during 1989–1990 in the northeast atlantic (48°n 21°w). *Marine Chemistry*, 71, 251–264.

Rosell-Melé, A., Eglinton, G., Pflaumann, U., & Sarnthein, M. (1995). Atlantic core-top calibration of the u37k index as a sea-surface palaeotemperature indicator. *Geochimica et Cosmochimica Acta*, 59(15), 3099 – 3107.

Rühlemann, C., Mulitza, S., Müller, P. J., Wefer, G., & Zahn, R. (1999). Warming of the tropical atlantic ocean and slowdown of thermohaline circulation during the last deglaciation. *Nature*, 402(6761), 511–514.

Sachs, J. P. (2007). Cooling of northwest atlantic slope waters during the holocene. *Geophysical Research Letters*, 34(3).

Salimbeni, H. & Deisenroth, M. P. (2017). Doubly stochastic variational inference for deep gaussian processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (pp. 4591–4602). Red Hook, NY, USA: Curran Associates Inc.

Sarnthein, M., Winn, K., Jung, S. J. A., Duplessy, J.-C., Labeyrie, L., Erlenkeuser, H., & Ganssen, G. (1994). Changes in east atlantic deepwater circulation over the last 30,000 years: Eight time slice reconstructions. *Paleoceanography*, 9(2), 209–267.

Sawada, K. & Handa, N. (1998). Variability of the path of the kuroshio ocean current over the past 25,000 years. *Nature*, 392, 592–595.

Sawada, K., Handa, N., Shiraiwa, Y., Danbara, A., & Montani, S. (1996). Long-chain alkenones and alkyl alkenoates in the coastal and pelagic sediments of the northwest north pacific, with special reference to the reconstruction of emiliania huxleyi and gephyrocapsa oceanica ratios. *Organic Geochemistry*, 24(8), 751 – 764.

Schefuss, E., Schouten, S., & Schneider, R. (2005). Climatic controls of central african hydrology during the last 20,000 years. *Nature*, 437, 1003–6.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44, 522–536.

Schouten, S., Hopmans, E. C., Schefuß, E., & Damsté], J. S. S. (2002). Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? *Earth and Planetary Science Letters*, 204(1), 265 – 274.

Schulz, H., Emeis, K.-C., Erlenkeuser, H., [von Rad], U., & Rolf, C. (2002). The toba volcanic event and interstadial/stadial climates at the marine isotopic stage 5 to 4 transition in the northern indian ocean. *Quaternary Research*, 57(1), 22 – 31.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2), 461–464.

Scott, G. H. (1963). Uniformitarianism, the uniformity of nature, and paleoecology. *New Zealand Journal of Geology and Geophysics*, 6(4), 510–527.

Seki, O., Bendle, J. A., Harada, N., Kobayashi, M., Sawada, K., Moossen, H., Inglis, G. N., Nagao, S., & Sakamoto, T. (2014). Assessment and calibration of tex86 paleothermometry in the sea of okhotsk and sub-polar north pacific region: Implications for paleoceanography. *Progress in Oceanography*, 126, 254 – 266. Biogeochemical and physical processes in the Sea of Okhotsk and the linkages to the Pacific Ocean.

Seki, O., Kawamura, K., Ikehara, M., Nakatsuka, T., & Oba, T. (2004). Variation of alkenone sea surface temperature in the sea of okhotsk over the last 85 kyrs. *Organic Geochemistry*, 35(3), 347 – 354. Selected papers from the Eleventh International Humic Substances Society Conference.

Seki, O., Sakamoto, T., Sakai, S., Schouten, S., Hopmans, E. C., Sinninghe Damste, J. S., & Pancost, R. D. (2009). Large changes in seasonal sea ice distribution and productivity in the sea of okhotsk during the deglaciations. *Geochemistry, Geophysics, Geosystems*, 10(10).

Sepúlveda, J., Pantoja, S., Hughen, K. A., Bertrand, S., Figueroa, D., León, T., Drenzek, N. J., & Lange, C. (2009). Late holocene sea-surface temperature and precipitation variability in northern patagonia, chile (jacaf fjord, 44°s). *Quaternary Research*, 72(3), 400 – 409.

Shackleton, N. J., Fairbanks, R. G., Chiu, T.-C., & Parrenin, F. (2004). Absolute calibration of the greenland time scale: implications for antarctic time scales and for Δ^{14}c. *Quaternary Science Reviews*, 23(14), 1513–1522.

Shackleton, N. J., Hall, M. A., & Edith, V. (2000). Phase relationships between millennial-scale events 64,000-24,000 years ago. *Paleoceanography*, 15.

Shevenell, A., Ingalls, A., Domack, E., & Kelly, C. (2011). Holocene southern ocean surface temperature variability west of the antarctic peninsula. *Nature*, 470, 250–4.

Shintani, T., Yamamoto, M., & Chen, M.-T. (2011). Paleoenvironmental changes in the northern south china sea over the past 28,000years: A study of tex86-derived sea surface temperatures and terrestrial biomarkers. *Journal of Asian Earth Sciences*, 40(6), 1221 – 1229. Quaternary Paleoclimate of the Western Pacific and East Asia: State of the Art and New Discovery.

Siegenthaler, U., Stocker, T. F., Monnin, E., Lüthi, D., Schwander, J., Stauffer, B., Raynaud, D., Barnola, J.-M., Fischer, H., Masson-Delmotte, V., & Jouzel, J. (2005). Stable carbon cycle - climate relationship during the late pleistocene. *Science*, 310(5752), 1313–1317.

Sikes, E., Farrington, J., & Keigwin, L. (1991). Use of the alkenone unsaturation ratio u37k to determine past sea surface temperatures: core-top sst calibrations and methodology considerations. *Earth and Planetary Science Letters*, 104(1), 36 – 47.

Sikes, E. L., Howard, W. R., Samson, C. R., Mahan, T. S., Robertson, L. G., & Volkman, J. K. (2009). Southern ocean seasonal temperature and subtropical front movement on the south tasman rise in the late quaternary. *Paleoceanography*, 24(2).

Sikes, E. L. & Keigwin, L. D. (1994). Equatorial atlantic sea surface temperature for the last 30 kyr: A comparison of u37 k′ , δ18o and foraminiferal assemblage temperature estimates. *Paleoceanography*, 9(1), 31–45.

Sikes, E. L. & Keigwin, L. D. (1996). A reexamination of northeast atlantic sea surface temperature and salinity over the last 16 kyr. *Paleoceanography*, 11(3), 327–342.

Sikes, E. L., Volkman, J. K., Robertson, L. G., & Pichon, J.-J. (1997). Alkenones and alkenes in surface waters and sediments of the southern ocean: Implications for paleotemperature estimation in polar regions. *Geochimica et Cosmochimica Acta*, 61(7), 1495 – 1505.

Sinninghe Damsté, J., Schouten, S., Hopmans, E., Van Duin, A., & Geenevasen, J. (2002). Crenarchaeol: The characteristic core glycerol dibiphytanyl glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota. *Journal of Lipid Research*, 43(10), 1641–1651.

Skinner, L. C., Elderfield, H., & Hall, M. (2013). *Phasing of Millennial Climate Events and Northeast Atlantic Deep-Water Temperature Change Since 50 Ka Bp*, (pp. 197–208). American Geophysical Union (AGU).

Skinner, L. C. & Shackleton, N. J. (2004). Rapid transient changes in northeast atlantic deep water ventilation age across termination i. *Paleoceanography*, 19(2).

Skinner, L. C. & Shackleton, N. J. (2005). An atlantic lead over pacific deep-water change across termination i: implications for the application of the marine isotope stage stratigraphy. *Quaternary Science Reviews*, 24(5), 571–580.

Skinner, L. C., Shackleton, N. J., & Elderfield, H. (2003). Millennial-scale variability of deep-water temperature and $\delta^{18}o_{dw}$ indicating deep-water source variations in the northeast atlantic, 0–34 cal. ka bp. *Geochemistry Geophysics Geosystems - GEOCHEM GEOPHYS GEOSYST*, 4.

Smith, M., De Deckker, P., Rogers, J., Brocks, J., Hope, J., Schmidt, S., Lopes dos Santos, R., & Schouten, S. (2013). Comparison of u37k′, tex86h and ldi temperature proxies for reconstruction of south-east australian ocean temperatures. *Organic Geochemistry*, 64, 94 – 104.

Sonzogni, C., Bard, E., & Rostek, F. (1998). Tropical sea-surface temperatures during the last glacial period: A view based on alkenones in indian ocean sediments. *Quaternary Science Reviews*, 17(12), 1185 – 1201.

Sonzogni, C., Bard, E., Rostek, F., Dollfus, D., Rosell-Melé, A., & Eglinton, G. (1997). Temperature and salinity effects on alkenone ratios measured in surface sediments from the indian ocean. *Quaternary Research*, 47(3), 344–355.

Sperling, M., Schmiedl, G., Hemleben, C., Emeis, K., Erlenkeuser, H., & Grootes, P. (2003). Black sea impact on the formation of eastern mediterranean sapropel s1? evidence from the marmara sea. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 190, 9 – 21.

Spratt, R. M. & Lisiecki, L. E. (2016). A late pleistocene sea level stack. *Climate of the Past*, 12(4), 1079–1092.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag New York.

Stern, J. V. & Lisiecki, L. E. (2014). Termination 1 timing in radiocarbon-dated regional benthic δ^{18}o stacks. *Paleoceanography*, 29(12), 1127–1142.

Stuart, A. & Ord, K. (1994). *Kendall's Advanced Theory of Statistics: Volume 1: Distribution Theory*. Wiley.

Tao, S., Xing, L., Luo, X., Wei, H., & Liu, Y. (2012). Alkenone distribution in surface sediments of the southern yellow sea and implications for the u 37k′thermometer. *Geo-marine Letters - GEO-MAR LETT*, 32, 61–71.

Ternois, Y., Kawamura, K., Ohkouchi, N., & Keigwin, L. (2000). Alkenone sea surface temperature in the okhotsk sea for the last 15 kyr. *GEOCHEMICAL JOURNAL*, 34(4), 283–293.

Terrell, G. R. & Scott, D. W. (1992). Variable kernel density estimation. *Ann. Statist.*, 20(3), 1236–1265.

Theodoridis, S. & Koutroumbas, K. (2008). *Pattern Recognition, Fourth Edition*. USA: Academic Press, Inc., 4th edition.

Tierney, J. E. & Tingley, M. P. (2014). A bayesian, spatially-varying calibration model for the tex86 proxy. *Geochimica et Cosmochimica Acta*, 127, 83–106.

Tierney, J. E. & Tingley, M. P. (2018). Bayspline: A new calibration for the alkenone paleothermometer. *Paleoceanography and Paleoclimatology*, 33(3), 281–301.

Tipping, M. E. & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 611–622.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In D. van Dyk & M. Welling (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research* (pp. 567–574). Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR.

Titsias, M. & Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research* (pp. 844–851). Chia Laguna Resort, Sardinia, Italy: PMLR.

Tjallingii, R., Claussen, M., Stuut, J.-B. W., Fohlmeister, J., Jahn, A., Bickert, T., Lamy, F., & Röhl, U. (2008). Coherent high- and low-latitude control of the northwest african hydrological balance. *Nature Geoscience*, 1, 670–676.

Tresp, V. (2001). Mixtures of gaussian processes. In *Advances in Neural Information Processing Systems 13* (pp. 654–660).: MIT Press.

Trommer, G., Siccha, M., [van der Meer], M. T., Schouten, S., Damsté], J. S. S., Schulz, H., Hemleben, C., & Kucera, M. (2009). Distribution of crenarchaeota tetraether membrane lipids in surface sediments from the red sea. *Organic Geochemistry*, 40(6), 724 – 731.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Verleye, T. (2011). *The late Quaternary palaeoenvironmental changes along the western South-American continental slope: a reconstruction based on dinoflagellate cysts and TEX86*. PhD thesis, Ghent University.

Villacampa-Calvo, C., Zaldivar, B., Garrido-Merchán, E. C., & Hernández-Lobato, D. (2020). Multi-class gaussian process classification with noisy inputs. arXiv:2001.10523v2 [stat.ML].

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260–269.

Waelbroeck, C., Skinner, L. C., Labeyrie, L., Duplessy, J.-C., Michel, E., Vazquez Riveiros, N., Gherardi, J.-M., & Dewilde, F. (2011). The timing of deglacial circulation changes in the atlantic. *Paleoceanography*, 26(3).

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wang, W. & Chen, X. (2016). The effects of estimation of heteroscedasticity on stochastic kriging. In *2016 Winter Simulation Conference (WSC)* (pp. 326–337).

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26, 359–372.

Wei, Y., Wang, J., Liu, J., Dong, L., Li, L., Wang, H., Wang, P., Zhao, M., & Zhang, C. L. (2011). Spatial variations in archaeal lipids of surface water and core-top sediments in the south china sea and their implications for paleoclimate studies. *Applied and Environmental Microbiology*, 77(21), 7479–7489.

Weldeab, S., Schneider, R. R., & Müller, P. (2007). Comparison of mg/ca- and alkenone-based sea surface temperature estimates in the fresh water–influenced gulf of guinea, eastern equatorial atlantic. *Geochemistry, Geophysics, Geosystems*, 8(5).

Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M., & Opper, M. (2018). Efficient gaussian process classification using polya-gamma data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33.

Williams, C. K. I. (1998). Computation with infinite neural networks. *Neural Computation*, 10(5), 1203–1216.

Williams, C. K. I. & Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1342–1351.

Wu, W., Tan, W., Zhou, L., Yang, H., & Xu, Y. (2012). Sea surface temperature variability in southern okinawa trough during last 2700 years. *Geophysical Research Letters*, 39(14).

Wuchter, C., Schouten, S., Coolen, M. J. L., & Sinninghe Damsté, J. S. (2004). Temperature-dependent variation in the distribution of tetraether membrane lipids of marine crenarchaeota: Implications for tex86 paleothermometry. *Paleoceanography*, 19(4).

Zarriess, M., Johnstone, H., Prange, M., Steph, S., Groeneveld, J., Mulitza, S., & Mackensen, A. (2011). Bipolar seesaw in the northeastern tropical atlantic during heinrich stadials. *Geophysical Research Letters*, 38(4).

Zarriess, M. & Mackensen, A. (2011). Testing the impact of seasonal phytodetritus deposition on δ^{13}c of epibenthic foraminifer cibicidoides wuellerstorfi: A 31,000 year high-resolution record from the northwest african continental slope. *Paleoceanography*, 26(2).

Zell, C., Kim, J.-H., Hollander, D., Lorenzoni, L., Baker, P., Silva, C. G., Nittrouer, C., & Damsté], J. S. S. (2014). Sources and distributions of branched and isoprenoid tetraether lipids on the amazon shelf and fan: Implications for the use of gdgt-based proxies in marine sediments. *Geochimica et Cosmochimica Acta*, 139, 293 – 312.

Zhao, M., Beveridge, N. A. S., Shackleton, N. J., Sarnthein, M., & Eglinton, G. (1995). Molecular stratigraphy of cores off northwest africa: Sea surface temperature history over the last 80 ka. *Paleoceanography*, 10(3), 661–675.

Zhao, M., Huang, C.-Y., Wang, C.-C., & Wei, G. (2006). A millennial-scale U37K' sea-surface temperature record from the South China Sea (8°N) over the last 150 kyr: Monsoon and sea-level influence. *Palaeogeography Palaeoclimatology Palaeoecology*, 236(1-2), 39–55.

Zhou, H., Hu, J., Spiro, B., Peng, P., & Tang, J. (2014). Glycerol dialkyl glycerol tetraethers in surficial coastal and open marine sediments around china: Indicators of sea surface temperature and effects of their sources. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 395, 114 – 121.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.