Model-based decision support with uncertain human-centric data

Jonathan M. Downing

Pembroke College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2019

Abstract

Both human and algorithmic decision making can be complex. To truly intertwine the two, algorithms need to understand the human decision making process and humans need to have transparent understanding of algorithmic decision analytics. In this thesis, we leverage Bayesian Gaussian processes to better understand people as well as offering a framework for people to better understand algorithms. Key to such transparency on both sides is the robust and principled reporting of uncertainty and careful consideration of preference: the latter being the foundational basis of human decision making. We consider pairwise preference modelling in which the strength of preference (e.g. strong or weak) can be seamlessly taken into account in the model, as well as confidence (uncertainty). We contribute to probabilistic functional regression by enabling the functional predictor space to be interpretable. We develop a principled, robust tool for understanding black-box algorithms, leading to a novel method for the interpretability of features in local domains. Finally, using our heteroscedastic ordinal regression model, we analyse the relationship between the features of occupations (such as their requirement for manual dexterity) and their expected share of US employment in 2030. Together, the approaches and algorithms we provide create a suite of models that increase our decision making capability with human-centric data.

Model-based decision support with uncertain human-centric data

Jonathan M. Downing

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

Acknowledgements

There is an uncountable number of people I would like to thank for positively influencing my life academically and socially over the last several years but in the necessity for brevity, I will only mention an important few.

Firstly, I would very much like to thank my supervisors Professor Stephen Roberts and Professor Michael Osborne for their unbounded support and guidance. For nurturing an environment where I was free to explore multiple academic interests and for being there for me when times were tough.

I am very grateful to Alex Brown for the unwavering and persistent support she has given me over the past 4 years. Always steadfast in her willingness to help in any way she thought would be beneficial.

To my lab collaborators, Justin Bewsher and Logan Graham it has been a pleasure working with you both and I am proud of the work we did together. I am also grateful for your friendship; you inspired and brought out the best in me. During the Future of Skill project I was privileged to work with Hasan Bakhshi (Nesta), Philippe Schneider (Nesta), Harry Armstrong (Nesta) and Mark Griffiths (previously Pearson).

Within the lab I am indebted to many for their caring camaraderie and stimulating ideas – I learned a lot from you guys. In particular Elmarie for lending her clarity, warmth and friendship throughout the whole process, Ali for encouraging me to be the best I could be, Rory for sharing his exquisite music taste with me during the early days, and Jack for having stimulating conversations, also a fellow Irishman that could actually understand what I said. A special mention for Mike at Eagle House reception who regularly welcomed me with tales of how it was in the 1940s and keeping my spirits high.

Outside of the lab, I split my time between socialising in Oxford and playing music in Northern Ireland. Within Oxford, I would like to thank Daniela and Saumya for their unwavering friendship and sharing their transcendental insights about life and the lives we live. I am indebted to my housemates, namely, Malte, Harriet, Yossi, and more recently Svenja for their support, solidarity and unrelenting fun throughout all these years. I also want to thank Olivia, Eveliina, Emma and Andrew for their friendship and who have made the last couple of years all the more fun.

I want to mention and thank my bandmates: Conchur, Christopher, Breandan and Micky, throughout the DPhil they have provided me with a beautiful musical

world to explore. We shared many wonderful experiences from playing at Glastonbury to South by South West in Austin, TX. I am grateful for their friendships, the good times we had and the music that we made.

I also want to thank Emily for her support over the period I was writing up, her help was invaluable.

A special thanks to my long-time friend Shea for taking the journey with me from St. Patrick's Grammar Armagh through undergraduate at Keble College, then both of us embarking on DPhils together. You have been there during my highest points and lowest points never faulting to help and support where you can.

Lastly, I would like to thank my Mum, Dad and Sister for always being there for me and spurring me on.

Abstract

Both human and algorithmic decision making can be complex. To truly intertwine the two, algorithms need to understand the human decision making process and humans need to have transparent understanding of algorithmic decision analytics. In this thesis, we leverage Bayesian Gaussian processes to better understand people as well as offering a framework for people to better understand algorithms. Key to such transparency on both sides is the robust and principled reporting of uncertainty and careful consideration of preference: the latter being the foundational basis of human decision making. We consider pairwise preference modelling in which the strength of preference (e.g. strong or weak) can be seamlessly taken into account in the model, as well as confidence (uncertainty). We contribute to probabilistic functional regression by enabling the functional predictor space to be interpretable. We develop a principled, robust tool for understanding black-box algorithms, leading to a novel method for the interpretability of features in local domains. Finally, using our heteroscedastic ordinal regression model, we analyse the relationship between the features of occupations (such as their requirement for manual dexterity) and their expected share of US employment in 2030. Together, the approaches and algorithms we provide create a suite of models that increase our decision making capability with human-centric data.

Contents

List of Figures

Special Terms

ARD Automatic Relevance Determination. 44, 49, 50, 52, 69, 72, 84, 85, 88, 90, 91, 136

ASEC Annual Social and Economic Supplement. 128

AUC Area Under the Curve. 5, 46, 47, 49, 50, 52–56, 62, 64, 66, 68, 152

BLS US Bureau of Labor Statistics. 124, 130, 146, 157, 158

CAM Continuously Additive Model. 75, 76

CCA Corpus Callosum. 89–91

CPS Current Population Survey. 128, 129

DNN Deep Neural Network. 4, 108, 110

DOT Dictionary of Occupational Titles. 126

DTI Diffusion Tensor Imaging. 4, 90, 91

EU European Union. 4, 96

FA Fractional Anisotropy. 89–91

FAM Functional Additive Model. 75, 87, 90

FAT Fat. 88, 90

FGAM Functional Generalized Additive Model. 5, 75, 76, 83–85, 88, 90, 91, 93, 182

FLM-Basis Functional Linear Basis Model. 87, 88, 90

FLM-PCA Functional Linear Functional Principal Component Model. 87, 90

FPC Functional Principal Component. 75, 87

FPCA Functional Principal Component Analysis. 87

FQM Functional Quadratic Model. 87, 90

FV Functional Kernel Model. 87, 88, 90

GCV Generalised Cross Validation. 88

GDPR General Data Protection Regulation. 4

GFLM Generalised Functional Linear Model. 87, 90

GP Gaussian Process. 1–8, 10, 14, 15, 20–22, 24–30, 32, 34, 41, 47, 49–52, 57, 59, 72, 73, 75–79, 82, 84, 85, 88, 92, 93, 102, 103, 115, 132, 133, 136, 139, 144, 181, 182, 185

GP-FGAM Gaussian Process Functional Generalized Additive Model. 8, 10, 76, 79, 82–88, 90–93, 182, 184

GP-HOR Gaussian Process Heteroscedastic Ordinal Regression. 7–10, 13, 58, 59, 64, 67, 68, 70, 116, 117, 119, 133–135, 145, 178, 179, 182, 183, 185

GP-OPL Gaussian Process Ordinal Preference Learning. 7, 8, 10, 13, 41, 44, 49, 50, 52–56, 60, 69, 70, 181

GP-OR Gaussian Process Ordinal Regression. 1, 3, 5, 8, 27, 36, 67, 68, 70, 182

GP-PL Gaussian Process Preference Learning. 3, 5, 8, 36, 37, 41, 48–50, 52–56, 69, 70, 181

HOR Heteroscedastic Ordinal Regression. 6, 36, 60, 62

IPUMS Integrated Public Use Microdata Series. 128, 129

KL Kullback–Leibler. 30, 65

L-BFGS Limited-memory Broyden–Fletcher–Goldfarb–Shannon. 143

LIME Local Interpretable Model-agnostic Explanations. 4, 96, 97, 108, 110

MCMC Markov Chain Monte Carlo. 27, 28

MD Mean Diffusivity. 89–91

MEG Magnetoencephalography. 184

MRI Magnetic Resonance Imaging. 184

NAICS North American Industry Classification System. 129

NIR Near Infra-Red Reflectance. 88

O*NET Occupational Information Network. 95, 120, 124–128, 132, 133, 138–140, 146, 158–166, 168–175, 179, 184, 185

OPL Ordinal Preference Learning. 36

PASAT Paced Auditory Serial Additional Test. 89

PIGEBaQ Principled Interpretability for Gradient Evaluation using Bayesian Quadrature. 8, 10, 72, 99, 100, 102–104, 107–115, 182–184

PRO Protein. 88, 90

QSAR Qualitative Structure Activity Relationships. 63

RankNet Rank Neural Network. 44, 45, 47, 49–52, 55, 56, 70, 181

RankSVM Rank Support Vector Machine. 44, 45, 47, 49–53, 55, 56, 70, 181

RCST Right Corticospinal. 89–91

ReLU Rectified Linear Unit. 45

RMSE Root Mean Square Error. 84, 85, 90, 91, 93, 109, 111, 112, 182

ROC receiver operating characteristic. 46, 64, 151–153

SE Squared Exponential. 23, 44, 62, 84, 85, 88, 90, 115, 136

SHAP SHapley Additive exPlanations. 4, 5, 108, 115

SNR Signal to Noise. 83–87, 93, 182

SOC Standard Occupation Classification. 124, 126, 129, 146, 149, 172, 173

SVM Support Vector Machine. 44

TEC Tecator. 88–92, 184

VUS Volume Under the Surface. 152, 153

WAT Water. 88, 90–92

1 Introduction

Contents

1.1 Motivation

Our world is becoming more complex. With each coming day, there are more people, more data created, and more decisions to be made impacting those people. Machine learning algorithms are now ubiquitous in our modern society, making decisions for us, from recommending your next movie to aiding the next scientific discovery; and automating tasks and occupations affecting millions of people across the world. This thesis explores improving the bi-directional understanding in human-computer interaction. Using the models developed in this thesis we investigate the future of skills needed to weather large-scale trends, including automation.

This thesis focuses on models involving *human-centric data*, which is defined as any data that is created by a human and is digitisable, as shown in Figure 1.1.

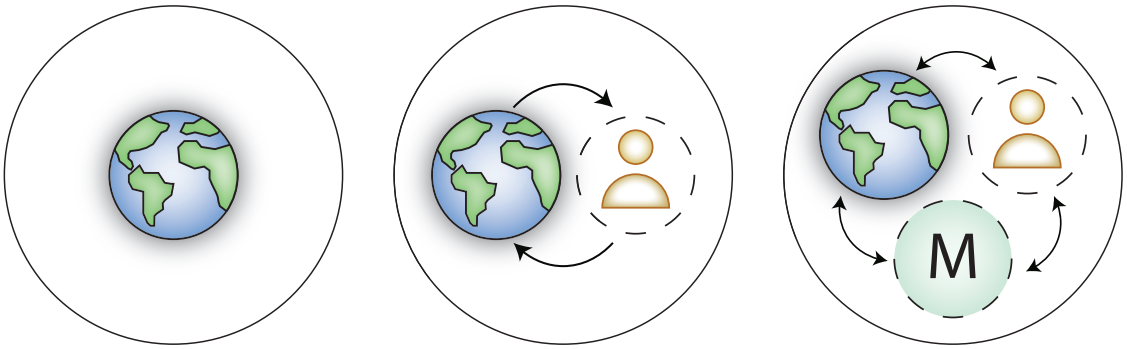


Figure 1.1: We note three stages of the world around us. The first being pre-human and all interactions are within and between the environment. The second being the interaction of humans and the environment. The third being the processing of the human and natural world by digitalised machines. *Human-centric data* in this instance is all information flowing from humans to machines.

Inherent in human-centric data is uncertainty, often due to lack of information. Probabilistic approaches are critical to ensure uncertainty is captured and these methods form the foundations of many machine learning algorithms. Uncertain quantities are described using distributions rather than brittle point estimates. This provides robust predictions whilst tempering overconfidence. There are many probabilistic models, but one which is predominant in the field is the Gaussian Process (GP) (C.E. Rasmussen and C. K. Williams 2006). Our interest lies in GPs, as they provide a principled way to model complex non-linear functions probabilistically. The GP is a Bayesian non-parametric functional prior defined by a mean and covariance function providing a flexible non-linear representation of data. GPs have been used to approach a range of machine learning problems, such as: time-series modelling, Bayesian optimisation and latent variable modelling (Bergstra et al. 2011; Roberts et al. 2013; Titsias and Lawrence n.d.; Snoek et al. 2012). Although much attention has been given to the GP there are still theoretical areas yet to be addressed. Three aspects that have received minimal attention are: pairwise preference and heteroscedastic extensions to ordinal models in GPs, functional regression with GPs, and model interpretability with GPs. There exist many applications that lie in conjunction with these theoretical GP areas, and one salient application that has only recently been explored is the Future of Employment

and Future of Skills in the economy. This thesis is focused on the bi-directional understanding between humans and machines using GPs in these areas.

Ordinal data models Ordinal models are integral to eliciting the preferences of people and represent the flow of information from a person to a machine. Conventionally, in an ordinal setting a number of ordered options are presented to a person, where each option represents the different degree of preference towards an instance (Chu and Ghahramani 2005b). We explore two types of data which are of particular interest in the ordinal setting: pairwise preferences and heteroscedastic ordinal labels.

Pairwise preference learning considers preference relations between two items, e.g. strawberry ice-cream is preferred to vanilla ice-cream. This is naturally a binary observation, i.e. preferred and not preferred. It is intuitively known that humans don't communicate exclusively through binary pairwise preference relations but can hold a wide range of strengths of preference. The current state-of-the-art Gaussian Process Preference Learning (GP-PL) model (Chu and Ghahramani 2005c) only considers binary pairwise preference relations. Extending this model to include the *strength of pairwise preferences* would therefore be beneficial.

Ordinal regression is a well known model which regresses over ordinal labels applied to items (Chu and Ghahramani 2005b). In the Bayesian setting it is commonly assumed that the observational noise is the same for all items. Just as before, it is intuitively known that the ordinal preference reported for one instance may be far more confident than a preference for an instance unknown to the person. The current state-of-the-art Gaussian Process Ordinal Regression (GP-OR) model (Chu and Ghahramani 2005b) does not consider heteroscedastic uncertainty on the labels. Developing this model to include the *heteroscedastic ordinal labels* would be useful, as again, it would more accurately capture human preferences.

Functional data models With both preference learning and ordinal regression, each instance is described by a finite set of features. With medical data such as Diffusion Tensor Imaging (DTI) data or other functional human-centric data the input domain is a partially observed function, with internal structure. This would

be poorly served by methods which do not capture structure between features. Work has been done in functional data models but has failed to utilise the non-linear probabilistic flexibility provided by a GP model (Morris 2015; J.-L. Wang et al. 2016). Therefore, exploring functional data models using GPs would be a worthwhile investigation to be made, as it would make it useful for processing human-centric functional data.

Model interpretability As many machine learning algorithms grow more complex (such as Deep Neural Networks (DNNs)) and take on a greater decision-making role in society, there is an explicit need for interpretability of those models. Globally there are calls for increased scrutiny and accountability with machine learning algorithms, especially when those algorithms are the basis of life-altering decisions. For example, an unemployed person fails to get a job because their algorithmically curated credit score is taken into account in the interview process and is too low. These injustices have caused demand, from both the public and academia, for a better understanding of how these decisions were made. It is increasingly important that we obtain transparency and clarity regarding what is happening 'under the hood' in these algorithms. These social implications of algorithm use, such as fairness and interpretability, are slowly being addressed by policy makers and through public conversations, such as recent legislation introduced in the European Union (EU), namely the General Data Protection Regulation (GDPR), which implies the need for explainability. In other jurisdictions similar legislation is coming into place (MIRON 2018). Attempts have been made to close the interpretability gap and they fall into model-agnostic and model-specific methods; work includes Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016b) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017). However, both fail to report the volatility of their estimates and, in the case of SHAP can be very sample inefficient. Therefore, using GPs for the model-agnostic interpretability problem would be a useful investigation to make.

Future of skills As a result of the rise of machines taking a larger role in society, such as automation in global industry, there is also a growing concern over the impact on employment. Frey and M. A. Osborne 2017, in their seminal paper, created a GP binary classification model of the probability of automation of a given occupation based on a set of skills. Building on this, it would be informative to policymakers to explore using heteroscedastic ordinal data from domain experts in a GP framework in order to capture the relationship between skills and future demand in the economy

1.2 Objectives

This thesis attempts to address four problems with human-centric data described in the introduction.

Firstly, we attempt to extend the capabilities of ordinal models in two ways using GPs, one being by introducing ordinal pairwise preferences and the other by introducing heteroscedastic ordinal labels. We investigate the possibility of extending the GP-PL model (Chu and Ghahramani 2005d) to include ordinal pairwise preference. Furthermore, we will compare our ordinal GP models against several baseline models using several synthetic datasets to assess the label Area Under the Curve (AUC) and Kendall's tau rank accuracy. For the second extension of ordinal models we investigate the possibility to extend GP-OR (Chu and Ghahramani 2005a) to include ordinal heteroscedastic labels. Following this we will test our model on a synthetic dataset to assess label AUC and Wasserstein distribution accuracy.

Secondly, we investigate if it is possible to develop a GP version of the Functional Generalized Additive Model (FGAM) model (Mathew W McLean et al. 2012). Our GP model will be tested on synthetic data to measure its predictive capability in terms of root mean square prediction error. Following this we will compare the proposed GP model against several baseline models using several real world functional human-centric data sets to determine our model's improvement. We investigate whether interpretable insights that strengthen our understanding of the natural phenomena can be garnered from our model.

Thirdly, we attempt to leverage GPs for model-agnostic interpretability of black-box models. More specifically, we aim to focus on the gradient interpretation of interpretability (Ribeiro et al. 2016b; Baehrens et al. 2010; Babibker and Goebel 2017). We will investigate if the average gradient and marginal gradient can be extracted.

Finally, as technological advance is causing major changes within the global economy, it is important to investigate the effect on the labour economy. (Frey and M. A. Osborne 2017) used a GP in their seminal work to shed light on the relationship between the automatability of jobs and the resulting changes in employment. We aim to use a GP Heteroscedastic Ordinal Regression (HOR) model in order to quantitatively study the relationship between the constituents of jobs, namely, skills and demand for those skills. From this GP model we will investigate if the importance of skills and their complementarities can be ascertained. We will also investigate whether we can use our model to find novel hypothetical occupations that will have high predicted demand.

1.3 Scope and limitations

This thesis is concerned with the bi-directionality of understanding between humans and machines using GPs; more specifically, through more nuanced ordinal models, functional data for regression, interpretability of black box models and the application of future of skills.

In all our work we assume that the output variable is drawn from a Gaussian distributed random variable, which is uni-modal.

The scope and limitations of each section are outlined below.

Ordinal models In both Gaussian Process Ordinal Preference Learning (GP-OPL) and Gaussian Process Heteroscedastic Ordinal Regression (GP-HOR) models we assume that there is an underlying utility function which fully describes a

person's preferences given an item. Therefore, we limit ourselves to only consider preferences that are explicitly transitive[1].

Our GP-HOR model is built on the assumption that the relationship between the score and the participant's confidence is linked by the observation noise of the score. The model is built on a 2-tuple observation. The first element is the participant's ordinal score and the second element is the self-reported ordinal confidence of that score; both must be present.

Functional regression We develop function-to-scalar problems under the assumption that functional trajectories are fully observed. Aiding inference and implementation, we make the assumption that our functional predictors are completely observed i.e. we assume noise-free observations of the underlying process. Function-to-function models are not explored. We further assume that the underlying functional surface is continuous, so this explicitly excludes discontinuities in functions, e.g. step functions.

Model interpretability In developing a model-agnostic GP model for interpretability we assume that the black-box function can be sampled multiple times and has a continuous input domain in every dimension and a continuous single-dimensional output co-domain. This limits us to investigating models that are reasonably well-behaved.

Future of skills We assume that the domain experts accurately ingest the current state of the labour market and can reliably map their understanding to a ternary ordinal scale of demand for each occupation queried. We also assume that the latent demand for each occupation is uni-modal and Gaussian distributed.

[1]Given $A > B$ and $B > C$, a transitive relation would be $A > C$

1.4 Outline

The following is a brief outline of each chapter for the benefit of the reader.

Chapter 2: Background theory The fundamental mathematical frameworks that are used in each chapter of this thesis are described. Probability theory is introduced and quickly followed by Gaussian Processes, a non-parametric model that is used extensively in this thesis. Specific Gaussian process models are explored, namely, GP regression, GP-OR and GP-PL.

Chapter 3: Ordinal models In this chapter we present two novel ordinal models, namely, GP-OPL and GP-HOR. For the GP-OPL model initially we motivate and explain *strength of pairwise preference* mathematically and from this develop the core theory using GPs. A number of synthetic datasets are employed to test the efficacy of the model and comparisons are made. The GP-HOR model is then motivated and developed. Comparisons are made with the homoscedastic case using a synthetic dataset. Directions for future work are described.

Chapter 4: Functional regression We present a novel approach to functional regression with GPs, namely Gaussian Process Functional Generalized Additive Model (GP-FGAM). The core theory is developed using GPs under the assumption of fully observed trajectories; we develop methods to perform inference and make predictions, in a simple, straightforward implementation. A number of synthetic examples are considered and we demonstrate the performance capability of GP-FGAM. GP-FGAM is also compared against comparison methods using real-world data.

Chapter 5: Model interpretability A novel approach to model-agnostic interpretability is developed and presented, namely, Principled Interpretability for Gradient Evaluation using Bayesian Quadrature (PIGEBaQ). Our method is validated against other comparison methods and using synthetic datasets derived from explicit functions. We close the chapter with an application of PIGEBaQ on a real-world dataset with an accompanying analysis.

Chapter 6: Future of skills This chapter presents novel work with a modified GP-HOR model to the application of economics and future studies. The introductory economic background material is introduced followed by a description of the data gathering workshops. The modified GP-HOR model is explained along with external datasets used in the model. A range of different results are presented covering aggregate future occupation demand predictions, granular skills demand predictions, predicted skill complementarities and new hypothetical occupations. Limitations and future work are presented.

Chapter 7: Conclusion We link together the key themes and findings from across the thesis, providing a narrative for the bigger picture, as well as reiterating the overall contributions of the thesis to the field of Machine Learning.

1.5 Contributions to this work

Below details the contributions made by myself and collaborators to this thesis.

Ordinal models

There are two novel GP models introduced in this chapter, namely, Gaussian Process Ordinal Preference Learning (GP-OPL) and Gaussian Process Heteroscedastic Ordinal Regression (GP-HOR). The theory for both GP-OPL and GP-HOR, along with the experimental results, was developed by myself. Stephen Roberts, Michael Osborne and Stephen Reece provided advice on general theory.

Functional Regression

The GP-FGAM model theory was joint work between myself and Justin Bewsher. The model was implemented by myself. Justin Bewsher identified the real world data while experiments were run jointly with me. Michael Osborne, Stephen Roberts and Jan-Peter Calliess provided advice on general theory.

Model interpretability

The theory for the PIGEBaQ model was developed primarily on my own with advice from Justin Bewsher, Logan Graham, Michael Osborne and Stephen Roberts. Justin Bewsher worked out the derivative for the Squared exponential kernel case. Logan Graham identified the real world data while all experiments were run by myself. Michael Osborne supplied problem motivation.

Future of Skills

This work was a collaboration with the educational charity Nesta and the publisher Pearson Plc., which resulted in a report published by Pearson Plc.:

Hasan Bakhshi, Jonathan M Downing, Michael A Osborne, and Philippe Schneider (2017). *The future of skills: employment in 2030*. Pearson

The theory for the modified GP-HOR model was developed by myself with guidance from Michael Osborne. All subsequent results from that model were produced by myself. Michael Osborne, Hasan Bakhshi, Philippe Schneider and Logan Graham wrote the report. Michael Osborne provided code for the aggregate plots (e.g. Figure 6.4). I also contributed by implementing the workshop active learning model and the live website interface used in the workshop. Logan Graham and I prepared the slides used in the live interface in the workshop. Justin Bewsher produced the trends analysis in Section 6.4.5. Hasan Bakhshi and Philippe Schneider at Nesta provided the economic commentary for the whole report. Michael Osborne and Hasan Bakhshi project managed the team. Harry Armstrong and Wendy L. Schultz ran the foresight workshops in the US (and UK). George Windsor, Juan Mateos-Garcia, Geoff Mulgan, Antonio Lima and Cath Sleeman at Nesta and Laurie Forcier, Amar Kumar, Janine Matho, Tom Steiner, Vikki Weston and other colleagues at Pearson provided feedback for the report. Mark Griffiths at Pearson played an important role in the inception of the study and provided valuable support throughout its implementation. Jessica Bland from Nesta also played a role guiding the design of the workshops during the early stages of the research.

2
Background theory

Contents

2.1 Introduction

The problems introduced in Chapter 1 are fundamentally supervised learning problems. In their simplest form, we observe both inputs and corresponding outputs and build models to predict the output for an unseen input, as illustrated in Figure 2.1. Also, all of the problems are inherently uncertain in their nature; this demands a principled framework to manage the flow of uncertainty. Bayesian probability theory is chosen as the principled framework to build our models with.

One mathematical model, in particular, that has been used throughout this

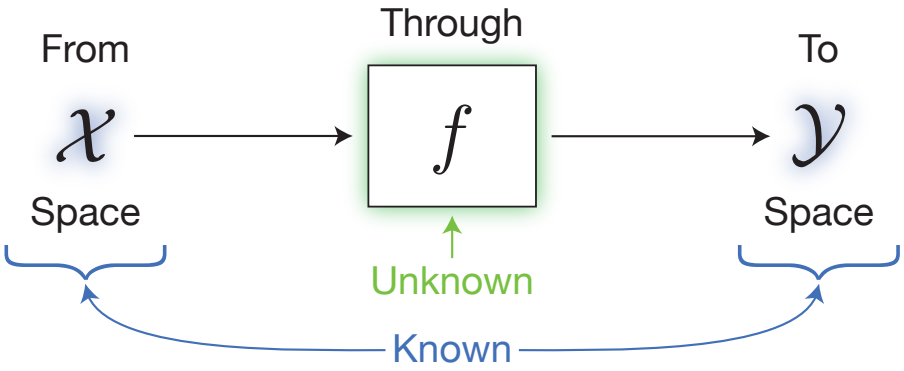


Figure 2.1: In the supervised learning problem we aim to infer the mapping between the input space $\mathcal{X}$ and output $\mathcal{Y}$ space by observing example pairs of the input and output.

thesis is the Gaussian Process (GP). A GP is a popular Bayesian non-parametric model which provides a principled and flexible method for incorporating uncertainty and is considered to have an infinite dimensional parameter space (Orbanz and Teh 2011). GPs are employed as flexible priors for latent variables that we aim to infer and correlations between points become more detailed as more data are observed (C. E. Rasmussen 2006a).

We will focus on three models built upon GPs, namely, GP regression, ordinal regression and preference learning. In GP regression the input variable is a multi-dimensional vector with a continuous scalar output variable. GP ordinal regression is similar to the continuous regression setting but the output variable is finite and ordinal. GP preference learning differs in that the input variable is a pair of multidimensional vectors and the output variable is binary. These three models are used extensively throughout this thesis and will be explained in greater depth.

2.1.1 Overview

The basis of all of our work begins with probability theory and Bayes' Theorem, this is explained in more detail in Section 2.2. GPs are introduced and explained in Section 2.3. This is followed by three derivative GP models, namely, GP regression, GP ordinal regression and GP preference learning in Section 2.4.

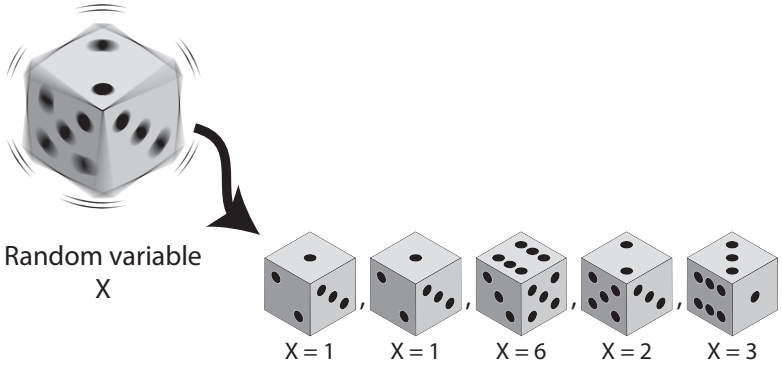


Figure 2.2: A random variable X isn't just free to take any random value. It is bounded by its domain $[1, \ldots, 6]$ and its probability density, $1/6$ for each face.

2.2 Probability theory

The problems set in the introduction chapter are inherently uncertain and in the very simple example of Figure 2.2 we see a die being thrown a number of times. If we knew the exact physical conditions of the die as it left the hand we would be able to predict the final landing position, but commonly we don't have access to that information. We are uncertain of the final outcome and we call a variable for which we don't have complete information a *random variable*.

As in the uncertainty in the problem of the die, our problems demand a principled framework. Bayesian probability theory provides a way to manage this uncertainty. Introducing a weaker notion, namely, plausibility of a statement as the progenitor of probability, Cox 1946 paved the way for a consistent Bayesian probability theory. Arnborg and Sjödin 2001 restate three requirements of any good theory of *plausibility of statements* from E.T. Jaynes' posthumous manuscript (Jaynes 2003), elucidating R.T. Cox's axioms (Cox 1946):

1. Divisibility and comparability – The plausibility of a statement is a real number and is dependent on information we have related to the statement.

2. Common sense – Plausibilities should vary sensibly with the assessment of plausibilities in the model.

3. Consistency – If the plausibility of a statement can be derived in two ways, the two results must be equal.

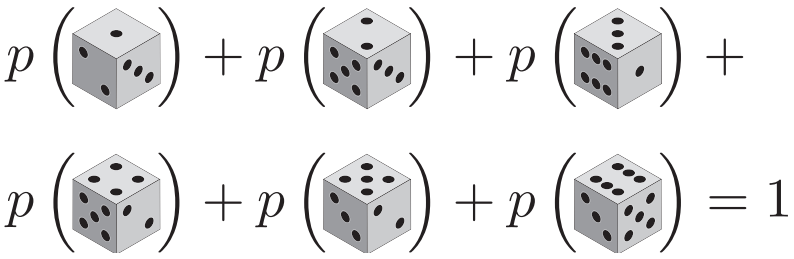


Figure 2.3: The probability of a one or two or ... or six must all sum to 1. It is with this simple principle that the sum of all the probabilities of every possible event must equal 1.

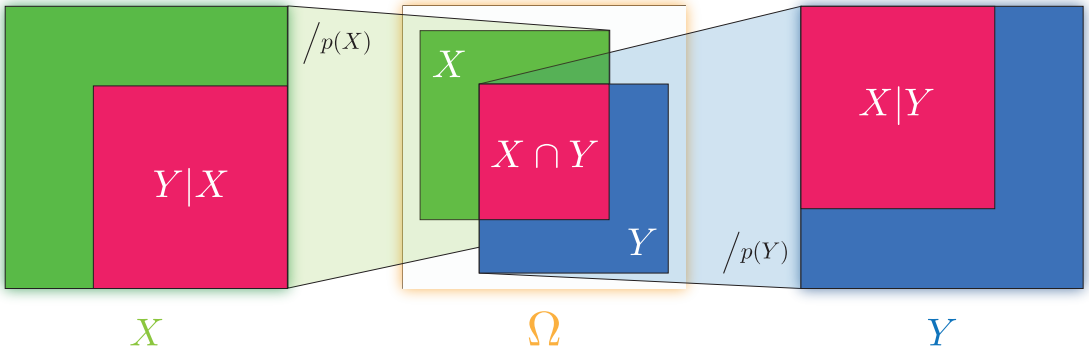


Figure 2.4: Bayes' Theorem: This is a seminal theorem that describes the probability of an event conditioned on another event.

As stated by Jaynes 2003, this leads to a number of results, the first being plausibility of a statement taking a functional form, one that maps the *statement* to the *plausibility*. Let us define this function as p : statement $\mapsto$ plausibility. The second is that certain truth is represented as $p\,(\text{certain truth}) = 1$ and certain falsehood as $p\,(\text{certain falsehood}) = 0$, thus probability lies between 0 and 1. The third is the *sum rule* of probability, that is the sum of all possible statements is equal to certain truth.

$$\sum_i p\,(\text{statement}_i) = 1 \tag{2.1}$$

Thinking back to our example of the die it is quite clear to see that (due to our grounding gravity) the die is certain to land and when it lands it is certain to take one of six positions. Therefore summing up the plausibility of each die value (as shown in Figure 2.3) is surely equal to certain truth.

Taking the time now to visualise in Figure 2.4 a simple case of two statements, X and Y which overlap within a space of all possible statements Ω. The overlap,

or intersection, between statements X and Y in the centre square is the joint probability between X and Y, or $p(X \cap Y) = p(X, Y)$. If we assume that the statement of Y becomes absolutely certain, i.e. the plausibility/probability becomes 1, we then scale the joint probability by $\frac{1}{p(Y)}$. This rescaled joint probability is now the probability that X is true conditional on Y being certainly true, i.e. $p(X|Y)$. This leads us to another important rule of probability, namely, the *product rule*:

$$p(X, Y) = p(X|Y)\, p(Y), \tag{2.2}$$

which can be recursively applied to many statements or as in our case variables:

$$p(X, Y, Z) = p(X|Y, Z)\, p(Y|Z)\, p(Z), \tag{2.3}$$

It is also order agnostic:

$$p(X, Y) = p(X|Y)\, p(Y) \tag{2.4}$$

$$= p(Y|X)\, p(X). \tag{2.5}$$

By equating each permutation this property naturally leads to *Bayes' Rule*:

$$p(X|Y) = \frac{p(Y|X)\, p(X)}{p(Y)}, \tag{2.6}$$

where $p(X|Y)$ is the posterior probability of X given Y, $p(Y|X)$ is the likelihood of Y given X, $p(X)$ is the prior probability of X and $p(Y)$ is commonly referred to as the evidence ($p(Y) = \int p(Y|X)\, p(X)\, dX$ by our sum rule).

Choosing a prior is an important decision to make to ensure that there is sufficient probability mass over all plausible statements. In the face of little information about the form of the distribution of the prior we seek principled ways of selecting a distribution. The principle of Maximum Entropy gives a method for calculating a distribution with knowledge only of the statement (sample) space, maximising our uncertainty of the distribution given the constraints. Assuming that our sample space has infinite support in n dimensions ($\mathbb{R}^n$), a specified mean μ and variance Σ our maximum entropy distribution will be Gaussian (Jaynes 2003):

$$p(X) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(\frac{1}{2}(X - \mu)^T \Sigma^{-1} (X - \mu)\right). \tag{2.7}$$

Also the Central Limit Theorem states that the sum of identically distributed independent variables (of any distribution that is sufficiently concentrated) in the limit of an infinite sum, is a Gaussian distribution. Therefore in systems where hidden variables have been inadvertently marginalised out in the process of observing a variable the resulting variable will be drawn from a Gaussian distribution.

Gaussian distribution identities

It's important to note the properties of Gaussian distributions that are useful and will be utilised in this thesis.

Sum rule The first being that a Gaussian distribution is a *stable distribution*. A stable distribution is one where the sum of two independent stable random variables results in another stable random variable of the same type. Therefore mathematically this makes Gaussian distributions applicable to many problems.

For example, $Z = X + Y$ where X, Y are independent Gaussian random variables, $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ then:

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y), \tag{2.8}$$

where $(\mu_X, \Sigma_X), (\mu_Y, \Sigma_Y)$ are the mean, variance pairs for X, Y, respectively.

Marginalisation The second property of Gaussian variables that is useful is marginalisation. Let us begin by describing two correlated Gaussian variables X and Y with joint probability $p(X, Y)$ described by:

$$p(X, Y) = \mathcal{N}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \middle| \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix},\right) \tag{2.9}$$

where $\Sigma_{XX}, \Sigma_{XY}, \Sigma_{YX}, \Sigma_{YY}$ are the variances and covariances of Gaussian variables X and Y.

When we marginalise out one of the variables, say X, we are summing up the probabilities of all different possible values X can take, effectively removing it from our joint probability. This can be written mathematically as:

$$p(Y) = \int p(X, Y) dX. \tag{2.10}$$

In the case of Gaussian distribution this conveniently results in:

$$p(Y) = \mathcal{N}(Y|\mu_Y, \Sigma_{YY}), \tag{2.11}$$

where Y is only dependent on it's mean μ_Y and the Σ_{YY} component of the joint covariance.

Conditioning In the case where we know that a variable, say Y, has been observed taking a specific value y we can condition X on $Y = y$. This is achieved by applying the product rule, described in Equation 2.2, to condition one Gaussian variable X on another Y. For clarity the probability of X conditioned on Y is:

$$p(X|Y) = \frac{p(X,Y)}{p(Y)}. \tag{2.12}$$

Substituting in our joint probability $p(X,Y)$ (Eq. 2.9) and marginalised probability $p(Y)$ (Eq. 2.11):

$$p(X|Y) = \mathcal{N}\left(\begin{bmatrix}X\\Y\end{bmatrix}\middle|\begin{bmatrix}\mu_X\\\mu_Y\end{bmatrix}, \begin{bmatrix}\Sigma_{XX} & \Sigma_{XY}\\\Sigma_{YX} & \Sigma_{YY}\end{bmatrix}\right) \Big/ \mathcal{N}(Y|\mu_Y, \Sigma_{YY}). \tag{2.13}$$

Given the product of two Gaussian distributions is itself a Gaussian distribution (potentially with a Gaussian scaling function) (Bromiley n.d.) we know that the division of two Gaussian distributions is also a Gaussian distribution. Therefore:

$$p(X|Y) \propto \mathcal{N}(X|\mu_{X|Y}, \Sigma_{X|Y}), \tag{2.14}$$

where $\mu_{X|Y}$ and $\Sigma_{X|Y}$ are both functions of X and Y.

Using the matrix inversion lemma (see Appendix A.3 CE Rasmussen and C. Williams 2006) and lots of tedious algebra manipulation one can expand Equation 2.13 and equate terms with Equation 2.14 to obtain $\mu_{X|Y}$ and $\Sigma_{X|Y}$:

$$\mu_{X|Y} = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y) \tag{2.15}$$

$$\Sigma_{X|Y} = \Sigma_{XX} - \Sigma_{XY}^T\Sigma_{YY}^{-1}\Sigma_{XY}. \tag{2.16}$$

In the next section we will introduce the very useful GP, which endows the Gaussian distribution with an index set, thus creating a stochastic process.

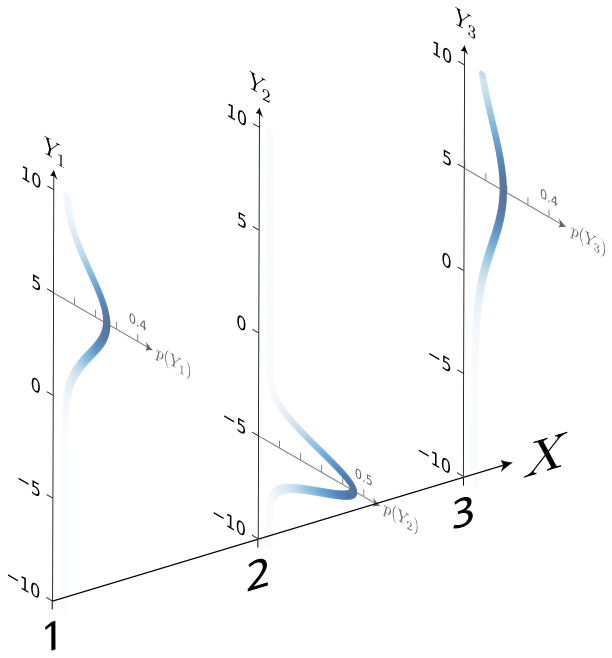


Figure 2.5: With three simple Gaussian distributions we can begin to see how as we add more Gaussian distributions indexed along the X we can build up a stochastic process. As it is described currently there would seem to be a need to explicitly define the mean and covariance for every distribution along the X axis. This is avoided by turning the mean and covariance into functions as described in the text. $\mu = [5; -5; 5]$ and $\Sigma = [1, 0.5, 0; 0.5, 0.8, 0.5; 0, 0.5, 1.5]$.

2.3 Gaussian Processes (GPs)

For the problems posed in the introduction chapter functional Bayesian priors are required and GP are an ideal candidate. Formally, a GP (CE Rasmussen and C. Williams 2006) is a probability distribution over functions $f : \mathcal{X} \mapsto \mathbb{R}$, such that the marginal distribution over the function values on any finite subset of $\mathcal{X}$ (such as X) is multivariate Gaussian.

To understand GPs a little better let us consider a simple example with three correlated Gaussian variables Y_1, Y_2 and Y_3 as illustrated in Figure 2.5. Each of these three variables are indexed by $X \in \mathcal{X}$, i.e. have a value corresponding to each random variable. In our case Y_1, Y_2 and Y_3 lie at $X = 1, 2$ and 3, respectively, where $\mathcal{X} \in \{1, 2, 3\}$. This provides a simple example of functional mapping from X to Y, where each index $X \in \mathcal{X}$ has a corresponding Gaussian random variable Y_X.

In the case where $\mathcal{X} \in \mathbb{R}$ or some other infinite set it is difficult to explicitly write down a complete joint probability between all variables. But we can write an expression of the mappings between a finite subset of $\mathcal{X}$ and the corresponding mean and covariance of the variables. The mean of one indexed random variable (e.g. Y_X) and the covariance between two indexed random variables (e.g. Y_X and $Y'_{X'}$) become functions, where $m : \mathcal{X} \mapsto \mathbb{R}$ and $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, respectively. We call $m(X)$ the *mean function* and $k(X, X')$ the *covariance function* or more commonly the *kernel function*. The kernel function is used to capture the structure between input variables. For example, we may wish to capture smoothness or we may know that our data are periodic and want that expressed in our function.

Bringing these together into a common notation a function f drawn from a GP with mean function $m(X)$ and kernel function $k(X, X')$ is written as:

$$f \sim \mathcal{GP}(m(X), k(X, X')) \tag{2.17}$$

It is also important to make clear that the Gaussian Process (GP) is a Bayesian non-parametric model (Ghahramani 2013), meaning that its expressivity will naturally adapt to that inherent in the data. GPs are considered to have a potentially infinite dimensional parameter space by virtue of this non-parametric nature (Orbanz and Teh 2011). This gives us an in-built resistance to *over-fitting* (learning patterns that do not generalise to unseen data): the model will not induce the flexibility required to give a near-perfect fit on training data unless the quality and quantity of data suggests that this fit will extend equally well to unobserved data. This desirable property is induced by the 'Occam's Razor' implicit within Bayesian reasoning (MacKay 2003).

The kernel function is used to construct the covariance matrix between all variables in the finite subset of $\mathcal{X}$, for example in the case of three variables indexed by X_1, X_2, X_3 the covariance matrix K is:

$$K = \begin{bmatrix} k(X_1, X_1) & k(X_2, X_1) & k(X_3, X_1) \\ k(X_1, X_2) & k(X_2, X_2) & k(X_3, X_2) \\ k(X_1, X_3) & k(X_2, X_3) & k(X_3, X_3) \end{bmatrix}. \tag{2.18}$$

Therefore a local change in the kernel function invokes a global change in the resulting GP whereas a local change in mean function invokes only a local change in the resulting GP. This illustrates why the kernel function is commonly the most interesting and important aspect of a GP, as opposed to the mean function.

It is critical that the resulting covariance matrix K be valid, meaning it is both symmetric and positive semi-definite for any possible subset of $\mathcal{X}$. To ensure that the covariance matrix K is symmetric we must satisfy the constraint: $k(X, X') = k(X', X)$. Ensuring matrix K is positive semi-definite is slightly more involved as the kernel function needs to satisfy the following constraint:

$$\int \int g(X)k(X, X')g(X')dXdX' \geq 0 \tag{2.19}$$

where $g(X)$ is any function.

So now that we have defined what constraints the kernel function must adhere to we can construct a valid covariance matrix. For multi-dimensional features let $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} \in \mathbb{R}^d$, where d is the dimensionality of the feature vector. A few examples of kernel functions and their corresponding function draws are presented below. Note that the kernel and function plots are 1-dimensional and $\boldsymbol{x}$ are therefore not bold ($x \in \mathbb{R}^1$).

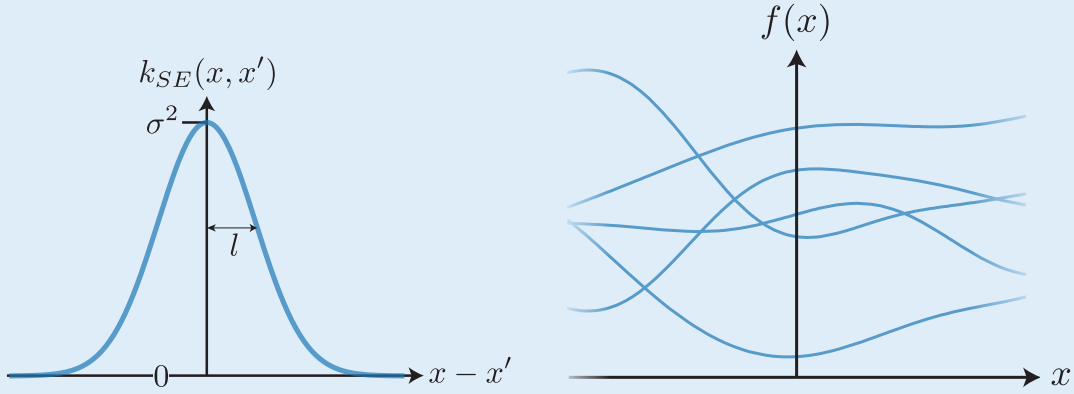
The very useful property of kernels is that we can add or multiply them in order to get composite behaviour (See C. E. Rasmussen 2006b for proofs that adding or multiplying kernels result in another valid kernel).

Squared Exponential (SE)

The most commonly used kernel is the Squared Exponential (SE) which has the form:

$$k_{SE}(\boldsymbol{x}, \boldsymbol{x'}) = \kappa^2 \exp\left(-\frac{1}{2}\frac{\|\boldsymbol{x} - \boldsymbol{x'}\|_2^2}{l^2}\right), \tag{2.20}$$

where κ^2 is the output variance, l is the lengthscale and $\|.\|_2$ is the l_2 norm.

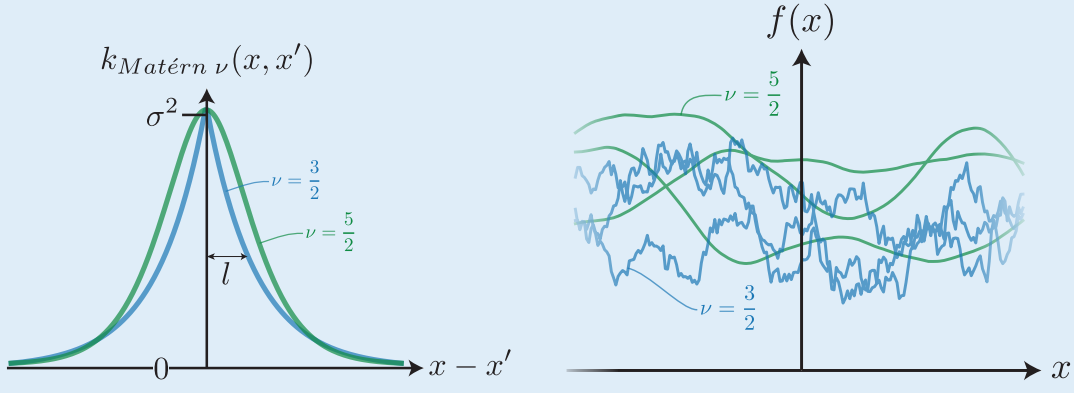


Matérn

Another common and very flexible kernel is the Matérn which has the form:

$$k_{Matérn\ \nu}(\boldsymbol{x}, \boldsymbol{x}') = \kappa^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2}{l}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2}{l}\right), \tag{2.21}$$

where κ^2 is the output variance, l is the lengthscale, Γ is the gamma function, K_ν is the modified Bessel function of the second kind and order ν. ν is commonly chosen to be $\frac{3}{2}$ or $\frac{5}{2}$ in practice.

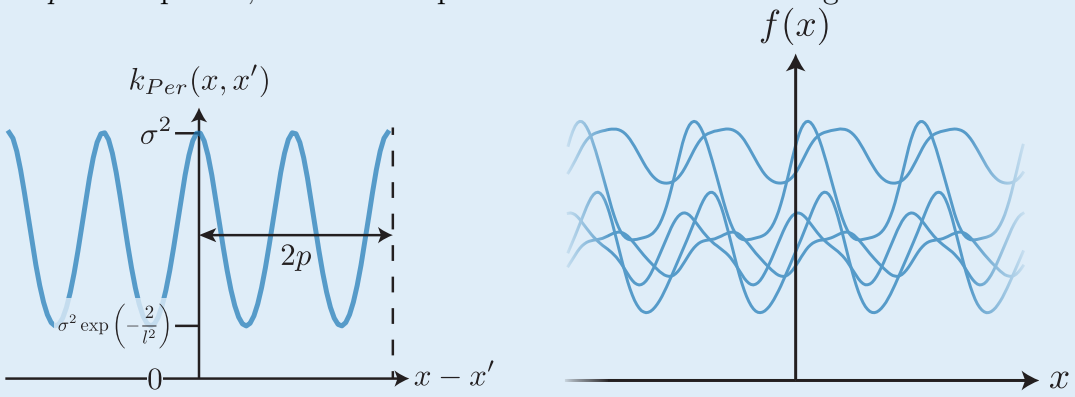


Periodic

The Periodic kernel is a commonly used kernel which offers global periodicity:

$$k_{Per}(\boldsymbol{x}, \boldsymbol{x}') = \kappa^2 \exp\left(-\frac{2\sin^2\left(\frac{\pi}{p}\|\boldsymbol{x} - \boldsymbol{x}'\|_2\right)}{l^2}\right), \tag{2.22}$$

where p is the period, κ^2 is the output variance and l is the lengthscale.

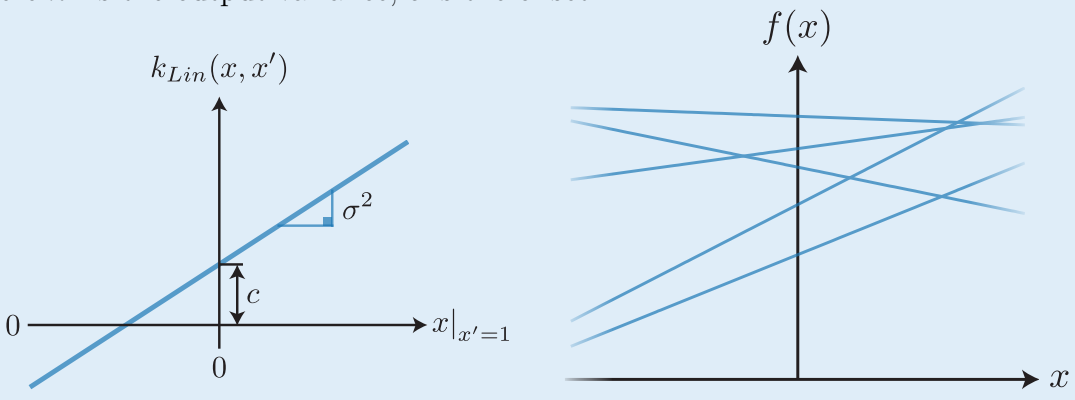


Linear

If the assumption is made that the functions are linear the following kernel is produced:

$$k_{Lin}(\boldsymbol{x}, \boldsymbol{x}') = \kappa^2 \boldsymbol{x}^T \boldsymbol{x}' + c, \tag{2.23}$$

where κ^2 is the output variance, c is the offset.



2.4 Gaussian Process models

We now provide background theory of the GP models which we will build on in subsequent chapters.

2.4.1 Gaussian Process regression

The essence of regression is being able to make a prediction from a model conditioned on observed data. Fortunately, the properties of Gaussian distributions, as described at the end of Section 2.2, make conditioning the GP on data analytically tractable.

Noise-free regression

We shall explore how to make predictions from a GP conditioned on noise-free data.

Let's assume we have a function $f \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$ and a vector of n noise-free observations $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^T$ corresponding the values of function f at points $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathcal{X}$.

Our goal is to calculate the distribution of f at a new unobserved point $\boldsymbol{x}^*$ conditional on past observations $\boldsymbol{f}$. In order to do this we condition $p(f^*)$ on the past observed distributions $p(\boldsymbol{f})$ as follows:

$$p(f^*|\boldsymbol{f}) = \frac{p(f^*, \boldsymbol{f})}{p(\boldsymbol{f})}. \tag{2.24}$$

Due to the self-conjugacy of the Gaussian distribution Equation 2.24 is analytic. Defining the joint distribution between f^* and $\boldsymbol{f}$ as:

$$p(f^*, \boldsymbol{f}) = \mathcal{N}\left(\begin{bmatrix} f^* \\ \boldsymbol{f} \end{bmatrix} \middle| \begin{bmatrix} \mu_{f^*} \\ \mu_{\boldsymbol{f}} \end{bmatrix}, \begin{bmatrix} K_{f^*f^*} & K_{f^*\boldsymbol{f}} \\ K_{\boldsymbol{f}f^*} & K_{\boldsymbol{ff}} \end{bmatrix}\right), \tag{2.25}$$

and the distribution of $\boldsymbol{f}$ as:

$$p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}|\mu_{\boldsymbol{f}}, K_{\boldsymbol{ff}}\right), \tag{2.26}$$

where:

$$\mu_{f^*} = m(\boldsymbol{x}^*) \tag{2.27}$$

$$\mu_{\boldsymbol{f}} = [m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_n)] \tag{2.28}$$

$$K_{f^*f^*} = k(\boldsymbol{x}^*, \boldsymbol{x}^*) \tag{2.29}$$

$$K_{f^*\boldsymbol{f}} = [k(\boldsymbol{x}^*, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}^*, \boldsymbol{x}_n)] \tag{2.30}$$

$$K_{\boldsymbol{f}f^*} = K_{f^*\boldsymbol{f}}^T \tag{2.31}$$

$$K_{\boldsymbol{f}\boldsymbol{f}} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \ldots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}. \tag{2.32}$$

Using the expressions above and with explicit reference to $\boldsymbol{x}^*$ Equation 2.24 can now be written as a single multi-variate Gaussian by incorporating Equations 2.25 & 2.26:

$$p(f^*|\boldsymbol{f}) = \frac{p(f^*, \boldsymbol{f})}{p(\boldsymbol{f})} \tag{2.33}$$

$$= \mathcal{N}\left(\begin{bmatrix} f^* \\ \boldsymbol{f} \end{bmatrix} \middle| \begin{bmatrix} \mu_{f^*} \\ \mu_{\boldsymbol{f}} \end{bmatrix}, \begin{bmatrix} K_{f^*f^*} & K_{f^*\boldsymbol{f}} \\ K_{\boldsymbol{f}f^*} & K_{\boldsymbol{f}\boldsymbol{f}} \end{bmatrix}\right) \Big/ \mathcal{N}\left(\boldsymbol{f}|\mu_{\boldsymbol{f}}, K_{\boldsymbol{f}\boldsymbol{f}}\right) \tag{2.34}$$

The calculations of this equation is described in Section 2.2 and results in a single Gaussian of the form:

$$p(f^*|\boldsymbol{f}) = \mathcal{N}\left(f_*|\, \mathbb{E}\left[f^*|\boldsymbol{f}\right], \mathbb{V}\mathrm{ar}\left[f^*|\boldsymbol{f}\right]\right), \tag{2.35}$$

where

$$\mathbb{E}\left[f^*|\boldsymbol{f}\right] = K_{f^*\boldsymbol{f}} K_{\boldsymbol{f}\boldsymbol{f}}^{-1} \boldsymbol{f} \tag{2.36}$$

$$\mathbb{V}\mathrm{ar}\left[f^*|\boldsymbol{f}\right] = K_{f^*f^*} - K_{f^*\boldsymbol{f}} K_{\boldsymbol{f}\boldsymbol{f}}^{-1} K_{\boldsymbol{f}f^*}. \tag{2.37}$$

GP regression with homoscedastic observational noise

Often the observations made are noisy. A common way of incorporating this observational uncertainty is to assume the observations are corrupted with additive Gaussian noise.

Let us assume we make independent noisy observations y of an underlying noise-free function $f \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$[1]. Given we assume additive Gaussian noise we can state:

$$y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon, \tag{2.38}$$

where $\epsilon \sim \mathcal{N}\left(0, \sigma_y^2\right)$ and σ_y^2 is the variance of the observational noise on y.

A Gaussian distribution, as stated before, is a stable distribution and the addition of two independent Gaussian distributions results in Gaussian distribution. Given a new unobserved point $\boldsymbol{x}^*$ we are interested in predicting our corresponding value y^*conditional on past observations $\boldsymbol{y}$, where $\boldsymbol{y} = [y^1, \ldots, y^n]$.

Using the rules of probability and the tractability of Gaussian distributions, our predictive equations, given a new unobserved point $\boldsymbol{x}^*$, are:

$$\mathbb{E}\left[y^* | \boldsymbol{x}^*, \boldsymbol{y}\right] = K_{f^*\boldsymbol{f}}(K_{\boldsymbol{f}\boldsymbol{f}} + \sigma_y^2 \mathbf{I})^{-1}\boldsymbol{f} \tag{2.39}$$

$$\mathbb{V}\mathrm{ar}\left[y^* | \boldsymbol{x}^*, \boldsymbol{y}\right] = K_{f^*f^*} - K_{f^*\boldsymbol{f}}(K_{\boldsymbol{f}\boldsymbol{f}} + \sigma_y^2 \mathbf{I})^{-1}K_{\boldsymbol{f}f^*}. \tag{2.40}$$

2.4.2 Gaussian Process Ordinal Regression

Ordinal regression is very similar to traditional regression with the key difference being the output is an ordinal variable not a continuous variable. This brings challenges, especially in the Bayesian GP domain where the output is naturally continuous. Therefore, a link function (Chu and Ghahramani 2005b) is employed in order to map between the continuous and ordinal domains.

This transforms the problem of inference from an analytical calculation to one that is not trivial due to the inherent non-Gaussian distribution the link (likelihood) function will induce. There are many methods used for inference ranging from Markov Chain Monte Carlo (MCMC) methods and variational approximation techniques to simpler techniques such as Laplace's approximation.

Laplace's approximation is a Gaussian approximation $q(\boldsymbol{f} \mid \mathcal{D})$ to the posterior $p(\boldsymbol{f} \mid \mathcal{D})$, whereby performing a second order Taylor expansion of $\log p(\boldsymbol{f} \mid \mathcal{D})$ results

[1] A zero mean function is assumed with no loss in generality.

in:

$$q(\boldsymbol{f} \mid \mathcal{D}) = \mathcal{N}(\boldsymbol{f}|\tilde{\boldsymbol{f}}, A^{-1}), \tag{2.41}$$

where $\tilde{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} p(\boldsymbol{f} \mid \mathcal{D})$ is the mode of the posterior and $A = -\nabla_{\boldsymbol{f}}\nabla_{\boldsymbol{f}} \log p(\boldsymbol{f} \mid \mathcal{D})|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}}$ (CE Rasmussen and C. Williams 2006). This method is cheap to compute and works well when the posterior $p(\boldsymbol{f} \mid \mathcal{D})$ is unimodal and approximately Gaussian but in cases where the posterior has heavily skewed this approximation fails to capture where the probability mass lies. E.g. in the case where $p(\boldsymbol{f} \mid \mathcal{D})$ is composed of step function link function and a Gaussian prior resulting in a truncated Gaussian:

$$p(\boldsymbol{f} \mid \mathcal{D}) = \begin{cases} \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{I}) & where \|\boldsymbol{f}\| \geq 0 \\ 0 & where \|\boldsymbol{f}\| < 0 \end{cases}, \tag{2.42}$$

where $\boldsymbol{I}$ is the identity matrix. Applying Laplace's approximation results in:

$$\tilde{\boldsymbol{f}} = \boldsymbol{0} \tag{2.43}$$

$$A = \boldsymbol{I}. \tag{2.44}$$

Therefore using Laplace's approximation in this instance resulted in the posterior reverting back to the zero-mean prior, completely failing to capture the true posterior distribution. Since our true posterior could take this form we will not be using Laplace's approximation.

MCMC methods such as Hybrid MCMC or Annealed Importance sampling are exact as the number of samples tends towards infinity but this can take a very long time to compute generally and due to the highly correlated nature of GPs makes the process even longer (Kuss and C. E. Rasmussen 2006).

Variational approximate inference (Opper and Archambeau 2008), as described later in this section, captures the probability mass of the posterior distribution more accurately compared to Laplace's approximation, whilst being far more computationally efficient compared to MCMC methods (Kuss and C. E. Rasmussen 2006).

Model definition As stated in (Chu and Ghahramani 2005b), let the training dataset of n observations be: $\mathcal{D} = \{(\boldsymbol{x}_i, y^{(i)})|i = 1, ..., n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, $y^{(i)} \in [1, ..., r]$, d is the dimensionality of the data, r is the total number of ordinal labels.

The noise-free latent function $f(\boldsymbol{x})$ has a GP prior:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \tag{2.45}$$

where $m(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x}')$ are the mean and kernel functions, respectively, of the latent function f.

We can describe the link function which maps the continuous domain of f to the ordinal domain of y :

$$y = \begin{cases} r & a_{r-1} < f(\boldsymbol{x}) \leq a_r \\ \vdots & \vdots \\ 1 & a_0 < f(\boldsymbol{x}) \leq a_1 \end{cases}, \tag{2.46}$$

where $[a_0, \ldots, a_r | a_i \in \mathbb{R}]$ are the latent threshold values.

Likelihood The likelihood of y wrt. f for labels is described as follows:

$$p(y|f(\boldsymbol{x})) = \begin{cases} 1 & a_y < f(\boldsymbol{x}) \leq a_{y+1} \\ 0 & \text{else} \end{cases} \tag{2.47}$$

We assume the observations are corrupted by additive Gaussian noise, where the observation label noise is independent (i.e. homoscedastic):

$$p(y|f(\boldsymbol{x})) = \begin{cases} 1 & a_y < f(\boldsymbol{x}) + \epsilon \leq a_{y+1} \\ 0 & \text{else} \end{cases}, \tag{2.48}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and σ^2 is the variance of the observational label noise.

Marginalising out the noise ϵ from $p(y|f(\boldsymbol{x}))$ results in:

$$p(y|f(\boldsymbol{x})) = \Phi\left(\frac{a_{y+1} - f(\boldsymbol{x})}{\sigma}\right) - \Phi\left(\frac{a_y - f(\boldsymbol{x})}{\sigma}\right), \tag{2.49}$$

where $\Phi(u) = \int_{-\infty}^{u} \mathcal{N}(\zeta|0, 1)\, d\zeta$.

Inference

For this thesis we employ the variational Gaussian approximation to infer the posterior distribution, as detailed in (Opper and Archambeau 2008). This assumes multivariate Gaussian distribution $q(\boldsymbol{f})$ can be used to approximate $p(\boldsymbol{f}|\mathcal{D})$, where:

$$q(\mathbf{f}) = \mathcal{N}\left(\mathbf{f}; \mathbf{K_{ff}}\boldsymbol{\alpha}, \mathbf{L}\mathbf{L}^T\right), \tag{2.50}$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]$, $\mathbf{K_{ff}}$ is the covariance matrix with elements $\mathbf{K}_{\mathbf{ff}}^{(i,j)} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The variational parameters $\boldsymbol{\alpha}$ are the variational mean parameters for the approximate posterior, $\mathbf{L}$ is square lower triangular matrix and the variational Cholesky parameters for the approximate posterior.

The optimal approximation is achieved when the Kullback–Leibler (KL) divergence[2] between $q(\boldsymbol{f})$ and $p(\boldsymbol{f}|\mathcal{D})$ is minimised and this is equivalent to minimising the variational free energy (evidence lower bound):

$$F(q, \boldsymbol{\theta}) = -\sum_i \mathbb{E}_{q(f_i)}\left[\log p(y_i \,|\, f_i)\right] - \frac{N}{2}\ln 2\pi e - \frac{1}{2}\ln|\mathbf{L}\mathbf{L}^T|, \tag{2.51}$$

where N is the number of data and $\boldsymbol{\theta}$ are the collection of variational parameters and model hyperparameters.

The expectation of the log likelihood $\mathbb{E}_{q(f_i)}\left[\log p(y_i \,|\, f_i)\right]$ takes the form:

$$\mathbb{E}_{q(f_i)}\left[\log p(y_i \,|\, f_i)\right] = \int_{-\infty}^{\infty} q(f_i) \log p(y_i \,|\, f_i) df_i \tag{2.52}$$

and which can be approximated using quadrature with n evaluations:

$$\int_{-\infty}^{\infty} q(f_i) \log p(y_i \,|\, f_i) df_i = \int_{-\infty}^{\infty} \omega(f_i) I(f_i) df_i \tag{2.53}$$

$$\approx \sum_{j=1}^{n} w^j I(f_i^j), \tag{2.54}$$

where weighting function $\omega(f_i) = q(f_i)$, integrand $I(f_i) = \log p(y_i \,|\, f^i)$, w^j are weights, f_i^j are latent values, $j \in [1, ..., n]$, n is the number of quadrature points. Given $\omega(f_i)$ is a Gaussian ($\propto \exp -x^2$) the optimal placement of f_i^j values are at

[2]Defined as $D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$

the roots of a Hermite polynomial H_n of order n and the weights w^j are given by (Abramowitz et al. 1965 Equation 25.4.46):

$$w^j = \frac{2^{n-1}n!\sqrt{\pi}}{n^2[H_{n-1}(f_i^j)]^2}. \tag{2.55}$$

This is called Gauss-Hermite quadrature and is optimal for functions I which can be described by a polynomial of degree $2n-1$ or below (Weisstein n.d.).

The Python libraries GPflow (Matthews, van der Wilk, Nickson, Keisuke. Fujii, et al. 2016) and Tensorflow (Martin Abadi et al. 2015) were employed for implementation. Within GPflow this integral is calculated using Gauss-Hermite quadrature (in all experiments the number of Gauss-Hermite points = 10).

For model selection, the variational free energy is minimised, with the addition of any hyperpriors:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} F(q, \boldsymbol{\theta}) + \log(p(\boldsymbol{\theta})), \tag{2.56}$$

the gradients of $F(q, \boldsymbol{\theta}) + \log(p(\boldsymbol{\theta}))$ are calculated automatically by the Tensorflow backend.

Prediction The posterior distribution of the label y at a new unobserved test instance $\boldsymbol{x}_*$ is calculated as follows. Firstly, as the posterior distribution of the latent function f given a new unobserved test instance $\boldsymbol{x}_*$ are calculated:

$$q(f^* \,|\, \boldsymbol{x}_*, \mathcal{D}) = \mathcal{N}\left(f^* \,|\, \mathbf{K}_{\mathbf{f}\star}^T \boldsymbol{\alpha},\, \mathbf{K}_{\star\star} - \mathbf{K}_{\mathbf{f}\star}^T \Sigma^{-1} \mathbf{K}_{\mathbf{f}\star}\right), \tag{2.57}$$

where $\mathbf{K}_{\mathbf{f}\star} = [k(\boldsymbol{x}_1, \boldsymbol{x}_*), k(\boldsymbol{x}_2, \boldsymbol{x}_*), ..., k(\boldsymbol{x}_n, \boldsymbol{x}_*)]^T$, $\mathbf{K}_{\star\star} = [k(\boldsymbol{x}_*, \boldsymbol{x}_*)]$ and $\Sigma = \mathbf{L}\mathbf{L}^T$

Secondly, from the above results the posterior distribution of the output ordinal label y^* given a new unobserved test instance $\boldsymbol{x}_*$ can be calculated:

$$q(y^*|\boldsymbol{x}_*, \mathcal{D}) = \int p(y^*|f^*) q\left(f^* \,|\, \boldsymbol{x}_*, \mathcal{D}\right) \mathrm{d}f^* \tag{2.58}$$

$$= \Phi\left(\frac{a_{y^*} - \mu_{f^*}}{\sqrt{\sigma^2 + \sigma_{f^*}^2}}\right) - \Phi\left(\frac{a_{y^*-1} - \mu_{f^*}}{\sqrt{\sigma^2 + \sigma_{f^*}^2}}\right). \tag{2.59}$$

2.4.3 Preference learning

Pairwise preference learning is more challenging than standard regression or classification as training data comprises of pairwise comparisons between items. This demands a more complex link function in order to map between two continuous outputs of the GP and the binary domain.

Again this transforms the problem of inference from an analytical calculation to one that is not trivial due to the inherent non-Gaussian distribution the link function will have. The inference of choice for this model is to take a variational approximation of the posterior distribution for the same reasons as in the Ordinal regression section above.

Model definition As stated in (Chu and Ghahramani 2005d), let us consider a set of n unique d dimensional items $\boldsymbol{x}_i \in \mathbb{R}^d$ denoted as $\mathcal{X} = \{\boldsymbol{x}_i : i = 1, ..., n\}$. We define an unobserved latent utility function $f : \mathcal{X} \mapsto \mathbb{R}$, which encodes pairwise preference relations and their corresponding strengths observed in the dataset. The latent function allows us to infer a total ordering of items from a small set of pairwise choices. Pairwise preference relations of the form $\boldsymbol{x}_u \succ \boldsymbol{x}_v$ are preserved in the latent function by $f(\boldsymbol{x}_u) > f(\boldsymbol{x}_v)$.

All together now consider a set of m observed pairwise preference relations on the items, denoted as:

$$\mathcal{Q} = \{\boldsymbol{x}_{u_k} \succ \boldsymbol{x}_{v_k} : k = 1, ..., m\}, \tag{2.60}$$

where u_k and v_k are the *preferred* and *not preferred* item indices, respectively.

We can describe the link function which maps the continuous domain of two items of f to the binary domain of y :

$$y = \begin{cases} 1 & f(\boldsymbol{x}_u) - f(\boldsymbol{x}_v) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.61}$$

Likelihood In the ideal noise free case:

$$p_{Ideal}(y|f(\boldsymbol{x}_u), f(\boldsymbol{x}_v)) = \begin{cases} 1 & f(\boldsymbol{x}_u) - f(\boldsymbol{x}_v) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.62}$$

We assume the observations are corrupted by additive Gaussian noise $\epsilon \sim \mathcal{N}(\epsilon; 0, \sigma_N^2)$, where σ_N^2 is the observational noise and the binary observation noise is independent (i.e. homoscedastic). ϵ is then marginalised out:

$$\begin{aligned} p(y|f(\boldsymbol{x}_u), f(\boldsymbol{x}_v)) &= \int_{-\infty}^{\infty} p_{Ideal}(y|f(\boldsymbol{x}_u) - f(\boldsymbol{x}_v) + \epsilon)\mathcal{N}(\epsilon; 0, \sigma_N^2)d\epsilon \\ &= \Phi\left(\frac{f(\boldsymbol{x}_u) - f(\boldsymbol{x}_v)}{\sigma_N}\right), \end{aligned} \tag{2.63}$$

where $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(\gamma; 0, 1)d\gamma$

Therefore, the total joint likelihood of all observations is:

$$p(\mathcal{D} \mid f) = \prod_i p(y_i \mid f(\boldsymbol{x}_{u_i}), f(\boldsymbol{x}_{v_i})) \tag{2.64}$$

Inference As the likelihood is non-conjugate to the prior distribution as in Section 2.4.2 we employ the same method of inference using the variational Gaussian approximation to infer the posterior distribution implemented in the Python library GPflow.

The main difference between inference in the ordinal instance and ordinal preference case is that the ordinal preference likelihood (Eq. 2.63) has two arguments as opposed to one in the ordinal instance case (Eq. 2.49). Therefore, in reference to Equation 2.51, within GPflow the term $\mathbb{E}_{q([f(\boldsymbol{x}_{u_i}), f(\boldsymbol{x}_{v_i})])} [\log p(y_i \mid f(\boldsymbol{x}_{u_i}), f(\boldsymbol{x}_{v_i}))]$ is calculated using a 2-D Gauss-Hermite quadrature (in all experiments the number of Gauss-Hermite points = 10 in both dimensions).

Prediction

There are two aims: the first is to calculate the distribution of the latent function f at a new unobserved instance, the second is to calculate the distribution of the strength label y given two unobserved items for comparison. Tackling the first

aim: given a test instance $\boldsymbol{x}_*$ the expectation and variance of the latent function f can be calculated as follows:

$$q(f^\star \,|\, \boldsymbol{x}_*, \mathcal{D}) = \mathcal{N}\left(f^\star \,|\, \mathbf{K}_{\mathbf{f}\star}^T \boldsymbol{\alpha},\, \mathbf{K}_{\star\star} - \mathbf{K}_{\mathbf{f}\star}^T \Sigma^{-1} \mathbf{K}_{\mathbf{f}\star}\right), \tag{2.65}$$

where $\mathbf{K}_{\mathbf{f}\star} = [k(\boldsymbol{x}_1, \boldsymbol{x}_*), k(\boldsymbol{x}_2, \boldsymbol{x}_*), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x}_*)]^T$.

Using this result we can achieve a prediction for the probability of binary preference label y_* given two unobserved items, $\boldsymbol{x}_*^A$ and $\boldsymbol{x}_*^B$:

$$q(y_*|\boldsymbol{x}_*^A, \boldsymbol{x}_*^B, \mathcal{D}) = \int\int p(y_*|f_*^A, f_*^B) q\left(\begin{bmatrix} f_*^A & f_*^B \end{bmatrix} \middle| \boldsymbol{x}_*^A, \boldsymbol{x}_*^B, \mathcal{D}\right) \mathrm{d}f_*^A \mathrm{d}f_*^B, \tag{2.66}$$

where the difference between f_*^A and f_*^B is the only relationship used in $p(y_*|f_*^A, f_*^B)$. Using the transform $\Delta f_\star = f_*^A - f_*^B$ we can combine this into a 1-D distribution:

$$q(y_*|\boldsymbol{x}_*^A, \boldsymbol{x}_*^B, \mathcal{D}) = \int p(y_*|\Delta f_\star) q\left(\Delta f_\star | \boldsymbol{x}_*^A, \boldsymbol{x}_*^B, \mathcal{D}\right) \mathrm{d}\Delta f_\star, \tag{2.67}$$

where $q\left(\Delta f_\star | \boldsymbol{x}_*^A, \boldsymbol{x}_*^B, \mathcal{D}\right) = \mathcal{N}\left(\Delta f_\star; \mu_{f_*^A} - \mu_{f_*^B}, \sigma^2_{f_*^A f_*^A} + \sigma^2_{f_*^B f_*^B} - 2\sigma^2_{f_*^A f_*^B}\right)$.

Deriving this further satisfies the second aim:

$$q(y_*|\boldsymbol{x}_*^A, \boldsymbol{x}_*^B, \mathcal{D}) = \Phi\left(\frac{\mu_{f_*^A} - \mu_{f_*^B}}{\sqrt{\sigma_N^2 + \sigma^2_{f_*^A} + \sigma^2_{f_*^B} - 2\sigma^2_{f_*^A f_*^B}}}\right). \tag{2.68}$$

2.5 Summary

These models are the foundation of the work presented in this thesis. GPs are used in every chapter. GP regression is built upon in Chapters 4 and 5. GP ordinal regression is built up in Chapter 3 which is applied in Chapter 6. GP preference learning is built up in Chapter 3.

3
Ordinal Models

Contents

3.1 Introduction

In this chapter we investigate two regression concepts centred on human preferences within a Bayesian framework, both of which are introduced in the Background Chapter 2. The first concept extends pairwise preference learning to include an

ordinal ranking of the pairwise preference relations themselves. The second concept expands ordinal regression to include discrete ordinal heteroscedastic noise. An overview and background of each concept, followed by detailed work-throughs, are presented in the sections below. The work detailed in this chapter is my own and makes a twofold contribution to knowledge, by:

1. Extending Gaussian Process Preference Learning (GP-PL) to include preference strength and preference indifference.

2. Extending Gaussian Process Ordinal Regression (GP-OR) to include finite discrete heteroscedastic noise.

3.1.1 Overview

This chapter is split into two sections: Ordinal Preference Learning (OPL) and Heteroscedastic Ordinal Regression (HOR). The first section exploring OPL is introduced in 3.2 with motivation for extending GP-PL given in 3.2.1. We define the problem setting and our model in Sections 3.2.2 and 3.2.3. In Section 3.2.4 we describe the synthetic experiments carried out in order to validate the model. This leads onto presenting the results of the synthetic experiments in Section 3.2.5.

The second section extending HOR is introduced in 3.3 and a motivation subsection is given in 3.3.1. We define the problem setting and our model in Sections 3.3.2. Synthetic experiments are carried out in Section 3.3.4, leading onto the results of those synthetic experiments in Section 3.3.4.

Finally, future work is described in Section 3.4 and the conclusion is presented in Section 3.5.

3.2 Ordinal preference learning

In this section we set out to extend the expressivity of pairwise preferences to include the concept of ordinal pairwise preferences. This captures granularity in human preferences that have previously been ignored. This concept has been explored in the multi criteria decision making literature (Malakooti 2000) described

as *strength of preference* and in Doyle's Prospects for Preference (Doyle 2004) where he discusses this as a more realistic representation of preference as opposed to treating all pairwise preferences equally.

We note two distinct advantages that extending GP-PL can offer:

1. Accounting for different strengths of preference, which users express naturally.
2. Decreasing the number of questions asked in order to learn more about a user's preferences.

3.2.1 Motivation

Standard pairwise preference learning (as described in the Background Chapter 2) establishes that all pairwise preferences are equal. For example, with ice-creams I may prefer strawberry over vanilla and strawberry over coffee flavours. Therefore, the relationship between vanilla and coffee is unknown but without any more information we would assume vanilla and coffee are quasi-equivalent preferences.

However, this approach presents a limited account of real life preference hierarchy, in which one comparison of two preferences is rarely similar in magnitude to another comparison. This can be illustrated when we dive a little deeper and ask how strongly we hold each preference. For example, I mildly like strawberry compared with vanilla but I love strawberry compared with coffee, which I detest. In this case the distinction is very clear as there is a weak and strong pairwise preference.

Treating pairwise preferences equally fails to capture the hierarchy of preference effectively and that pairwise preferences are inherently unequal.

3.2.2 Definition of ordinal preferences

Preference strength assumes one preference comparison can be more important than another. Two ordinal strength s states and indifference state will be assumed (with no loss of generality), where 0, 1 and 2 represent indifference, weak and strong preference, respectively.

As stated before in the background chapter the degree of preference for an item is defined using the latent utility function f. The problem of preference

learning is framed as a comparison of pairwise preferences. Uncertainty in the user's pairwise preference and strength indicators are implicit in the user's response and these are represented in our model as a distribution over latent functions and learned from the data.

When presenting the user with two options, we further ask them whether their preference is 'strong' or 'weak'. Depending on the application this can reflect either their strength in their choice (for example, "u^{th} item is definitely much better than v^{th} item") or their uncertainty in the choice ("Given the information I have, I think u^{th} item is better than v^{th} item"). We focus on the former, namely, preference strength.

Consider a set of n unique d dimensional items $\boldsymbol{x}_i \in \mathbb{R}^d$ denoted as $\mathcal{X} = \{\boldsymbol{x}_i : i = 1, ..., n\}$. We define an unobserved latent utility function $f : \mathcal{X} \mapsto \mathbb{R}$, which encodes pairwise preference relations and their corresponding strengths observed in the dataset. The latent function allows us to infer a total ordering of items from a small set of pairwise choices. Pairwise preference relations of the form $\boldsymbol{x}_u \succ \boldsymbol{x}_v$ are preserved in the latent function by $f(\boldsymbol{x}_u) > f(\boldsymbol{x}_v)$. The strength of pairwise preference $\boldsymbol{x}_u \succ \boldsymbol{x}_v$ is represented by the magnitude of the difference between $f(\boldsymbol{x}_u)$ and $f(\boldsymbol{x}_v)$, the greater the strength the larger the deviation.

Formalising this we introduce the difference function:

$$g(\boldsymbol{x}_u, \boldsymbol{x}_v) = f(\boldsymbol{x}_u) - f(\boldsymbol{x}_v), \tag{3.1}$$

where $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. This generalises the relationship between the two items. Therefore, the one-level strength pairwise preference correspondence becomes: $\boldsymbol{x}_u \succ \boldsymbol{x}_v \to g(\boldsymbol{x}_u, \boldsymbol{x}_v) > 0$.

While maintaining transitivity the strength of a preference can be measured by the relative magnitude of the function f, i.e. preference strength is then the absolute value of $g(\boldsymbol{x}_u, \boldsymbol{x}_v)$. Again without loss of generality we restrict our analysis to two main strength classes (Weak and Strong):

$$\begin{aligned} 0 < |g(\boldsymbol{x}_u, \boldsymbol{x}_v)| \leq t_b \quad & \text{Weak } (s = 1) \\ t_b < |g(\boldsymbol{x}_u, \boldsymbol{x}_v)| \leq \infty \quad & \text{Strong } (s = 2), \end{aligned} \tag{3.2}$$

where t_b is the weakest strong preference difference OR the strongest weak preference difference and the sign of $g(\boldsymbol{x}_u, \boldsymbol{x}_v)$ dictates either *preferred* or *not preferred*. This leads to four possible outcomes: *strongly not preferred*, *weakly not preferred*, *weakly preferred*, *strongly preferred*.

The strength label of a pairwise preference is defined as $s \in \mathbb{Z}^+$, where the maximum strength is defined as s_{max}, e.g. in the $s_{max} = 2$ setting there are two strength classes, a strong pairwise preference defined as $s = 2$ and the weak equivalent $s = 1$. When indifference is included this can be thought of as $s = 0$. We add a subscript s to the $\succ$ operator to indicate the strength level of the pairwise preference. E.g. $\boldsymbol{x}_u \succ_s \boldsymbol{x}_v$ is the u^{th} item is preferred to the v^{th} item with strength s.

Generalising this to any number of strength levels (with and without indifference) we define a function y mapping the continuous domain of g to an ordinal scale with r labels:

$$y(\boldsymbol{x}_u, \boldsymbol{x}_v) = \begin{cases} r & a_{r-1} < g(\boldsymbol{x}_u, \boldsymbol{x}_v) \leq a_r \\ \vdots & \vdots \\ 1 & a_0 < g(\boldsymbol{x}_u, \boldsymbol{x}_v) \leq a_1 \end{cases}, \tag{3.3}$$

where $[a_0, \ldots, a_r | a_i \in \mathbb{R}]$ are the latent threshold values which generalise the terms in Equation 3.2.

As the ordinal scale includes the pairwise preference relation and preference strength the total number of ordinal labels r is defined as:

$$r = \begin{cases} 2s_{max} & \text{without indifference} \\ 2s_{max} + 1 & \text{with indifference} \end{cases} \tag{3.4}$$

The mappings between the the specific pairwise strength comparison and y are detailed in Table 3.1.

To ensure that these symmetries are maintained in the latent domain of f we apply the following constraints:

Without indifference	**With indifference**
$y = \begin{cases} r & \boldsymbol{x}_u \succ_s \boldsymbol{x}_v \\ \vdots & \vdots \\ \frac{r}{2}+1 & \boldsymbol{x}_u \succ_1 \boldsymbol{x}_v \\ \frac{r}{2} & \boldsymbol{x}_u \prec_1 \boldsymbol{x}_v \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_u \prec_s \boldsymbol{x}_v \end{cases}$	$y = \begin{cases} r & \boldsymbol{x}_u \succ_s \boldsymbol{x}_v \\ \vdots & \vdots \\ \frac{r+1}{2}+1 & \boldsymbol{x}_u \succ_1 \boldsymbol{x}_v \\ \frac{r+1}{2} & \boldsymbol{x}_u \sim \boldsymbol{x}_v \\ \frac{r+1}{2}-1 & \boldsymbol{x}_u \prec_1 \boldsymbol{x}_v \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_u \prec_s \boldsymbol{x}_v \end{cases}$

Table 3.1: Mapping between pairwise preference relations ($\mathcal{Q}$) with strength ($\mathcal{S}$) and their corresponding ordinal labels y. r represents the maximum ordinal label. When the indifference label is included r is odd otherwise r is even.

Without indifference:

1. $a_0 = -\infty$
2. $a_r = +\infty$
3. $a_{\frac{r}{2}} = 0$
4. $a_i = -a_j | i > \frac{r}{2} \wedge j < \frac{r}{2}$
5. $a_{i+1} > a_i | i > \frac{r}{2}$

With indifference:

1. $a_0 = -\infty$
2. $a_r = +\infty$
3. $a_i = -a_j | i > \frac{r}{2} \wedge j < \frac{r}{2}$
4. $a_{i+1} > a_i | i > \frac{r}{2}$

As illustrated in Table 3.2 this can be described as the pairwise preference equivalent of ordinal classification/regression. Also it is worth noting with this notation one-level strength pairwise preference can be conceived as a type of binary classification, as shown in Table 3.2. The Cartesian product of $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ forms a single instance along with a binary label $[0, 1]$ if $\boldsymbol{x}_v \succ \boldsymbol{x}_u$ or $\boldsymbol{x}_u \succ \boldsymbol{x}_v$, respectively. To our knowledge this is the first time ordinal regression has been combined with pairwise preference learning.

All together now consider a set of m observed pairwise preference relations on the items, denoted as:

$$\mathcal{Q} = \{\boldsymbol{x}_{u_k} \succ_{s_k} \boldsymbol{x}_{v_k} : k = 1, ..., m\}, \tag{3.5}$$

where u_k and v_k are the *preferred* and *not preferred* item indices, respectively, and

a set of m strength indicators, denoted as:

$$\mathcal{S} = \{s_k : k = 1, ..., m\}. \tag{3.6}$$

Together the pairwise preference relations $\mathcal{Q}$ and binary strength $\mathcal{S}$ are combined in a tuple $\mathcal{D}$, defined as:

$$\mathcal{D} = (\mathcal{Q}, \mathcal{S}). \tag{3.7}$$

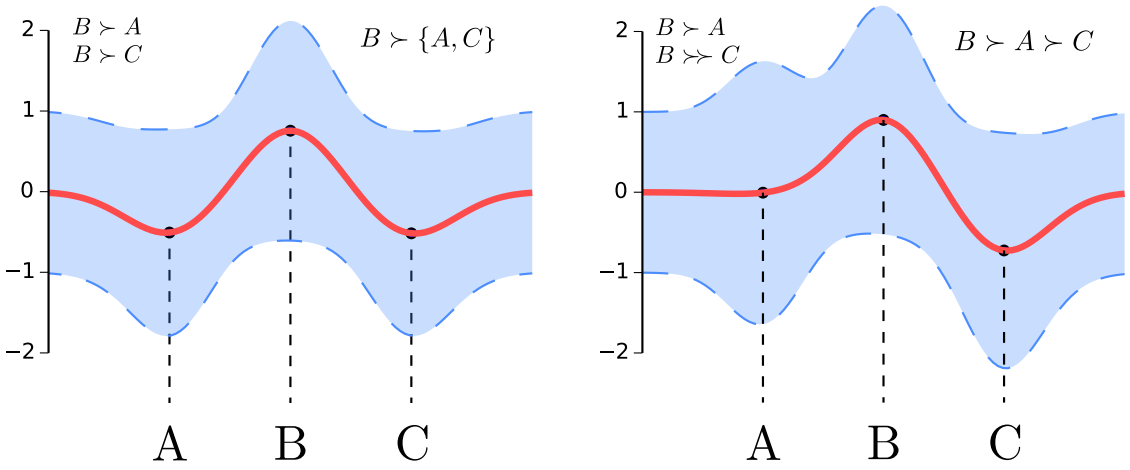
where each member of $\mathcal{Q}$ has a corresponding member in $\mathcal{S}$.

	Classification	Preference relations
Binary	$y(\boldsymbol{x}) = \begin{cases} 1 & g(\boldsymbol{x}) \geq 0 \\ 0 & else \end{cases}$	$y(\boldsymbol{x}_u, \boldsymbol{x}_v) = \begin{cases} 1 & g(\boldsymbol{x}_u, \boldsymbol{x}_v) \geq 0 \\ 0 & else \end{cases}$
Ordinal	$y(\boldsymbol{x}) = \begin{cases} r & a_{r-1} < g(\boldsymbol{x}) \leq a_r \\ \vdots & \vdots \\ 1 & a_0 < g(\boldsymbol{x}) \leq a_1 \end{cases}$	$y(\boldsymbol{x}_u, \boldsymbol{x}_v) = \begin{cases} r & a_{r-1} < g(\boldsymbol{x}_u, \boldsymbol{x}_v) \leq a_r \\ \vdots & \vdots \\ 1 & a_0 < g(\boldsymbol{x}_u, \boldsymbol{x}_v) \leq a_1 \end{cases}$

Table 3.2: "Completing the square": Within the machine learning literature binary classification, Ordinal classification/regression and (binary) preference learning have been extensively researched as stated before. Ordinal preference learning (the function within the red box) is a natural extension of standard preference learning and ordinal regression, and is understudied.

3.2.3 Gaussian Process Ordinal Preference Learning Model

Given the items $\mathcal{X}$ and strength pairwise preference observations $\mathcal{D}$ we develop a Bayesian model to infer the pairwise strength labels for unseen items. As a basis we start from (Chu and Ghahramani 2005c) GP-PL model, as described in the Background Chapter 2. The prior over the latent function f is a Gaussian Process (GP) with mean function $m(x)$ and kernel function $k(x, x')$. We assume a zero mean function with no loss of generality. In the following sections the ordinal preference likelihood will be defined, the inference method will be described along with some information on calculating predictions and optimising hyperparameters.



(a) Preference relations are equal: $B \succ A$ and $B \succ C$ resulting in the final ranking of $B \succ \{A, C\}$.

(b) Preference relations including strength: $B \succ A$ and $B \succ\succ C$ resulting in the final ranking of $B \succ A \succ C$.

Figure 3.1: Simple toy example with three items and two comparisons are made; the true ranking is $B \succ A \succ C$. The red line represents the expectation of the latent function and the shaded region represents the area within one standard deviation from the expectation. Incorporating preference strength enables the distinction between A and C to be made.

Likelihood

Our likelihood generalises (Chu and Ghahramani 2005c) pairwise preference likelihood to any ordinal preference relation. In the ideal noise free case:

$$p_{Ideal}(y|f(\boldsymbol{x}_u), f(\boldsymbol{x}_v)) = \begin{cases} 1 & a_{y-1} < f(\boldsymbol{x}_u) - f(\boldsymbol{x}_v) \leq a_y \\ 0 & \text{otherwise} \end{cases}, \tag{3.8}$$

where y is the output strength label describing both preference comparison and strength in a single ordinal scale, and $[a_0, \ldots, a_r | a_i \in \mathbb{R}]$ are the latent value preference thresholds.

At this stage it's also important to note different types of uncertainty, namely, aleatoric and epistemic uncertainty. Aleatoric uncertainty, also known as statistical uncertainty, captures noise inherent in the observations. This is in contrast to epistemic uncertainty which captures the uncertainty of whether or not the model is correct (Kiureghian and Ditlevsen 2009). As aleatoric uncertainty can exist in

all observations it is assumed there is zero mean Gaussian distributed additive noise ε with variance $\sigma_N^2/2$ on each preference comparison. In our case epistemic uncertainty is captured in the choice of kernel present within the prior.

It should be noted that there is a subtle distinction between preference strength and preference uncertainty, preference strength as defined earlier is the expected deviation between compared items, whereas preference uncertainty is the magnitude of σ_N a variable with no dependence on the input domain. Therefore, strength is an observation and uncertainty is a latent variable to be inferred. The additive noise ε can be marginalised out:

$$\begin{aligned} p(y|f(\boldsymbol{x}_u), f(\boldsymbol{x}_v)) &= \int_{-\infty}^{\infty} p_{Ideal}(y|f(\boldsymbol{x}_u)+\varepsilon, f(\boldsymbol{x}_v)+\varepsilon)\mathcal{N}(\varepsilon; 0, \sigma_N^2/2)d\varepsilon \\ &= \Phi(z_1^r) - \Phi(z_2^r), \end{aligned} \tag{3.9}$$

where $z_1^r = \frac{a_r - f(\boldsymbol{x}_u)+f(\boldsymbol{x}_v)}{\sigma_N}$, $z_2^r = \frac{a_{r-1} - f(\boldsymbol{x}_u)+f(\boldsymbol{x}_v)}{\sigma_N}$ and $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(\gamma; 0, 1)d\gamma$

Therefore, the total joint likelihood of all observations is:

$$p(\mathcal{D} \mid f) = \prod_i p(y_i \mid f(\boldsymbol{x}_{u_i}), f(\boldsymbol{x}_{v_i})) \tag{3.10}$$

Inference

As the likelihood has the same form as in Sections 2.4.2 and 2.4.3 we employ the same method of inference using the variational Gaussian approximation to infer the posterior distribution implemented in the Python library GPflow.

Prediction

Again prediction is very similar to Section 2.4.3 apart from the output variable y taking an ordinal value as opposed to a binary value. This is encoded in the likelihood (Eq. 3.9) as an additional $\Phi(.)$ term. Recalculating Equation 2.67 with the likelihood from Equation 3.9 results in the posterior distribution for new observed items A and B:

$$q(y_*|\boldsymbol{x}_*^A, \boldsymbol{x}_*^B, \mathcal{D}) = \Phi\left(\frac{a_{y_*} - (\mu_{f_*^A} - \mu_{f_*^B})}{\sqrt{\sigma_N^2 + \sigma_{f_*^A}^2 + \sigma_{f_*^B}^2 - 2\sigma_{f_*^A f_*^B}^2}}\right) \tag{3.11}$$

$$- \Phi\left(\frac{a_{y_*-1} - (\mu_{f_*^A} - \mu_{f_*^B})}{\sqrt{\sigma_N^2 + \sigma_{f_*^A}^2 + \sigma_{f_*^B}^2 - 2\sigma_{f_*^A f_*^B}^2}}\right). \tag{3.12}$$

3.2.4 Synthetic experiments

The goal of our algorithm is to maximise the accuracy of the test labels y and underlying ranking of the latent function f. A range of experiments were undertaken: the first being a comparison between (Chu and Ghahramani 2005c) (which is the equivalent of binary preference learning), a neural network approach, namely, Rank Neural Network (RankNet) (Burges et al. 2005), a linear Support Vector Machine (SVM) method, namely, Rank Support Vector Machine (RankSVM) (Joachims 2002), and our proposed Gaussian Process Ordinal Preference Learning (GP-OPL) method with a range of different strength labels. The second set of experiments focused on how the accuracy of our model varied as the ratio of strong labelled preferences varied from 0 to 1. The third set of experiments varied the size of the training synthetic dataset. All three sets of experiments are explained in greater detail in the following section.

In all the experiments we use an isotropic and anisotropic Gaussian kernel. The isotropic Gaussian kernel, also known as a Squared Exponential (SE) kernel, is defined as:

$$k_{SE}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa^2 \exp\left(-\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}{2l^2}\right), \tag{3.13}$$

where κ^2 is the signal variance and l is the characteristic lengthscale.

The anisotropic Gaussian kernel, also known as an Automatic Relevance Determination (ARD) kernel Neal 1995, is similarly defined as:

$$k_{ARD}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa^2 \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \mathrm{diag}(\boldsymbol{v})^{-2}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right), \tag{3.14}$$

where $\boldsymbol{v} \in \mathbb{R}^d$ is a vector of characteristic lengthscales corresponding to each dimension of an instance.

The RankSVM comparisons were made using the Python-wrapped library *dlib* (D. E. King 2009), which is written in C++ and includes the linear SVM model, RankSVM. RankSVM has one hyperparameter C which is a regularisation coefficient balancing the error on training data and norm of weights. RankSVM was trained on all of the training set. Grid search was used to optimise the hyperparameter C by maximising the Kendall's tau rank metric between the scores of each training value and the true scores from the dataset.

The RankNet comparisons were made using a Keras based Python implementation of (Burges et al. 2005). The implementation was taken from (Alcorn 2019) and comprises three densely connected hidden layers followed by a final single unit with a linear activation function. The three hidden layers have $[128, 64, 32]$ units respectively and Rectified Linear Unit (ReLU) activation functions were used in all hidden layers. Apart from the architecture the three main hyperparameters of this model were: batch size, number of epochs and learning rate. After experimentation the batch size was set to 8 and the number of epochs set to 10 for all experiments. A grid search between e^{-8} and e^{0} was used to optimise the learning rate l_r by maximising the Kendall's tau rank metric between the scores of each training value and the true scores from the dataset.

Synthetic ordinal preference datasets

As was done in past papers (Chu and Ghahramani 2005d) four well known regression datasets were adapted for the purposes of testing strength within pairwise preferences, summarised in Table 3.3. Where each dataset had the form $(\boldsymbol{x}_i, f_i^{True}) \in \mathcal{D}^{True}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $f_i^{True} \in \mathbb{R}$. All datasets were taken from (Dua and Graff 2017). All data items ($\boldsymbol{x}_i$) were linearly scaled to the unit hypercube $[0, 1]^d$.

All possible pairwise comparisons were calculated by creating an upper triangular matrix, U, of differences between the ground truth targets f^{True}, described by $q_{ij} = f_i^{True} - f_j^{True}\ \forall j \geq i$. It follows that preference pair i is preferable to preference pair j when $q_{ij} > 0$ and the pairwise preference relation set $\mathcal{Q}$ can be constructed accordingly.

	Dimensionality (d)	Training (m_{Train})	Test(m_{Test})
Boston Housing	12	200	800
Machine CPU	6	200	800
Pyrimidines	26	200	800
Triazines	60	200	800

Table 3.3: Summary of datasets: where d is the dimensionality of the feature vector, m_{Train} and m_{Test} are the number of training and test preference pairs, respectively.

Using the empirical distribution of the absolute values from matrix U a range of quantiles were calculated. The resulting quantile thresholds are used to partition pairwise comparisons into strength sets $\mathcal{S}$. The choice of quantiles varied depending on the experiment but fall into two categories: one-level strength ($s_{max} = 1$) resulting in a single quantile threshold, and two-level strength ($s_{max} = 2$) resulting in two quantile thresholds. This enables unique disjoint pairwise comparisons to be sampled for each strength set separately. For all experiments pairwise comparisons with an absolute differences in the 0 to $\frac{1}{7}$ quantile were deemed to be *indifferent*. If the strength label of *indifferent* wasn't being observed in an experiment those pairwise comparisons were removed from the synthetic dataset for that experiment.

From the corresponding relation and strength sets, $\mathcal{Q}$ and $\mathcal{S}$, respectively, ordinal labels y^{Test} can be constructed using the mapping described in Table 3.1.

Each strength set was randomly split into train and test subsets with a ratio corresponding to the overall training size m_{Train} and testing size m_{Test}, as shown in Table 3.3.

For all experiments there were $N = 50$ iterations of each experiment.

Metrics

In all the experiments the predictions were tested with two metrics, namely, Receiver operating characteristic (ROC) Area Under the Curve (AUC) and the Kendall's tau rank correlation coefficient. ROC AUC is used to measure the accuracy of the predicted labels of test pairwise items. The AUC was chosen as it captures both the true and false positive rates of the prediction, which is superior to only

calculating the binary accuracy of the predictions. As AUC is inherently a metric of binary classification, in order to assess the accuracy of ordinal label predictions we compute the AUC for each label in a *one-vs-all* approach as mentioned in (Waegeman et al. 2008) and take the average:

$$\text{Ordinal AUC} = \frac{1}{r}\sum_{i=1}^{r}\text{AUC}(\hat{y}_i^{Test}, \hat{y}_i^{Model}), \tag{3.15}$$

where:

$$\hat{y}_i^{Test} = \left[\mathbb{I}\left(y_1^{Test} = i\right), ..., \mathbb{I}\left(y_{m_{Test}}^{Test} = i\right)\right] \tag{3.16}$$

$$\hat{y}_i^{Model} = \left[p\left(y_1^{Model} = i\right), ..., p\left(y_{m_{Test}}^{Model} = i\right)\right], \tag{3.17}$$

and y_j^{True} and y_j^{Model} are the test label and model predicted label, respectively for the j^{th} test preference pairs.

Kendall's tau rank correlation coefficient is used to measure the accuracy of the predicted overall rank of the latent surface f given test pairwise items. It should be noted that the values of the latent surface f have no absolute meaning, only relative ordinal meaning, a metric was chosen that captures rank, namely, Kendall's tau rank correlation. This is done by comparing the ground truth targets f^{True} with the latent model variable/score f^{Model}. It should be noted that both f^{True} and f^{Model} are over single items $\boldsymbol{x}$.

The Kendall's tau ranges from $[-1, 1]$, where 1 signifies that the test rank and the predicted rank are the same and -1 signifies where all pairs of test and predicted rank are discordant. Kendall's tau is calculated as follows:

$$\tau(\boldsymbol{f}^{True}, \boldsymbol{f}^{Model}) = \frac{n_c - n_d}{n(n-1)/2}, \tag{3.18}$$

where $\boldsymbol{f}^{True} = [f^{True}(\boldsymbol{x}_1), \ldots, f^{True}(\boldsymbol{x}_n)]$, $\boldsymbol{f}^{Model} = [f^{Model}(\boldsymbol{x}_1), \ldots, f^{Model}(\boldsymbol{x}_n)]$, n is the number of items, n_c and n_d are the number of concordant and discordant pairs, respectively, between the ground truth target and latent model variable/score ranks.

For models that produce a point estimated latent model score the Kendall's tau will also be a point estimate (i.e. RankSVM and RankNet). In the case of the probabilistic GP models the latent model variable is a distribution; therefore, the Kendall's tau is also a distribution, which has to be sampled.

Strength ratios							
$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{4}{9}$	$\frac{5}{9}$	$\frac{6}{9}$	$\frac{7}{9}$	$\frac{8}{9}$

Table 3.4: Summary of the different strength ratios tested, where the quantile thresholds are calculated: [indifference, strength ratio + (1 − strength ratio) ∗ indifference] and indifference = $\frac{1}{7}$.

Experiment set 1: Accuracy comparisons The aim of these experiments were to assess each model's ability to accurately predict the correct pairwise comparison, correct total ranking and corresponding strength label (where appropriate). The two quantile thresholds used were: [indifference, strength ratio + (1 − strength ratio) ∗ indifference] where strength ratio = $\frac{1}{2}$ and indifference = $\frac{1}{7}$.

Experiment set 2: Varying strength ratio In the two strength case of *weak* and *strong* it is important to investigate what the ratio of these two labels will have on the accuracy. We tested a range of different ratios of weak and strong preferences, as shown in Table 3.4. For each strength ratio both indifference and no indifference labels were tested. As preference strength isn't a feature of the comparative methods they weren't included.

Experiment set 3: Varying training set size This used the same datasets and comparisons as the accuracy comparisons in Experiment set 1 but investigated the effect of different training set sizes. We vary the number of preferences observed (equal number of strong and weak preferences used). The different training set sizes are summarised in Table 3.5.

3.2.5 Results and Discussion

Below we present and discuss our experimental results.

Experiment set 1: Accuracy comparisons

The first set of experiments compared the accuracy of predicting pairwise preferences of the ordinal preference model against Chu and Ghahramani's method GP-PL,

	Proportion of training set from experiment one							
	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{4}{9}$	$\frac{5}{9}$	$\frac{6}{9}$	$\frac{7}{9}$	$\frac{8}{9}$
Boston Housing	23	45	67	89	112	134	156	178
Machine CPU	23	45	67	89	112	134	156	178
Pyrimidines	23	45	67	89	112	134	156	178
Triazines	23	45	67	89	112	134	156	178

Table 3.5: Summary of training pairwise preferences where the size of the set is varied proportional to the training set size of Experiment set one ($m_{Train} = 200$ for all datasets).

RankNet and RankSVM.

In three out of four datasets incorporating ordinal preference significantly decreased the error rate of predicting pairwise preferences compared to the Chu and Ghahramani's methodGP-PL, RankNet and RankSVM.

AUC Comparison results using the AUC metric are presented in Figure 3.2. In all datasets the accuracy of Chu's GP-PL model labels were the highest. In the Boston Housing, Machine CPU and Pyrimidines datasets GP-PL labels performed with an AUC upwards of 0.96, within the Triazines dataset upwards of 0.8. The inclusion of two level strength in the GP-OPL model resulted in it consistently being the next best performer when predicting labels. Interestingly, the AUC values of the GP-OPL model when indifference was included were relatively lower and in some cases (Machine CPU dataset) the worst performing. The other comparison methods, namely, RankNet and RankSVM normally performed better than the GP-OPL model with indifference and one-level strength but worse than all other models. Between the comparison methods RankNet generally performed better than RankSVM but had a much wider spread of results, regularly exhibiting outliers where the AUC = 0. ARD has minimal effect on the accuracy of labels for all GP models.

Kendall's tau Comparison results using Kendall's tau metric are presented in Figure 3.3. In three datasets, namely, Boston Housing, Pyrimidines and Triazines the GP-OPL model with two-level strength performed statistically (95% confidence of

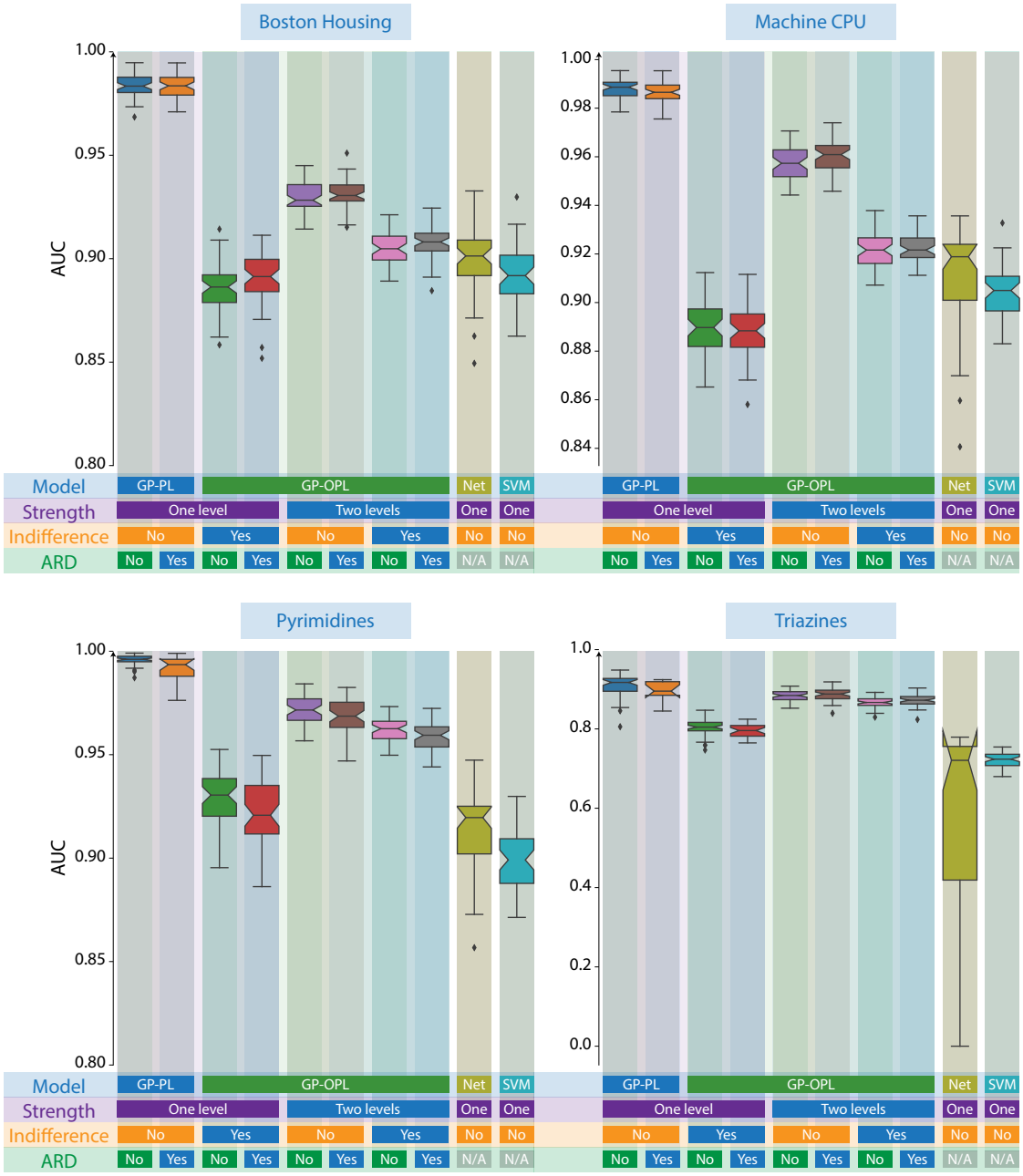


Figure 3.2: AUC accuracy boxplots for Experiment set 1. AUC range from $[0, 1]$, where 0, 1/2, and 1 are perfectly incorrect, random, perfectly correct ordinal classification, respectively. See text for details. For the Boston Housing and Pyrimidines datasets RankNet had 4 and 6 outliers, respectively, below the x-axis.

the median) better than GP-PL, RankNet and RankSVM. For the Pyramidines and Triazines datasets there were clear increases in ranking accuracy for GP-OPL when indifference was included compared when it was not. Interestingly, in the Machine CPU dataset the ranking accuracy for all GP models was less than RankNet. ARD

had a positive effect on GP models ranking accuracy within the Boston Housing and Triazines datasets; it was particularly pronounced in the Triazines dataset. Again RankNet performed better than RankSVM but had much wider spread of results.

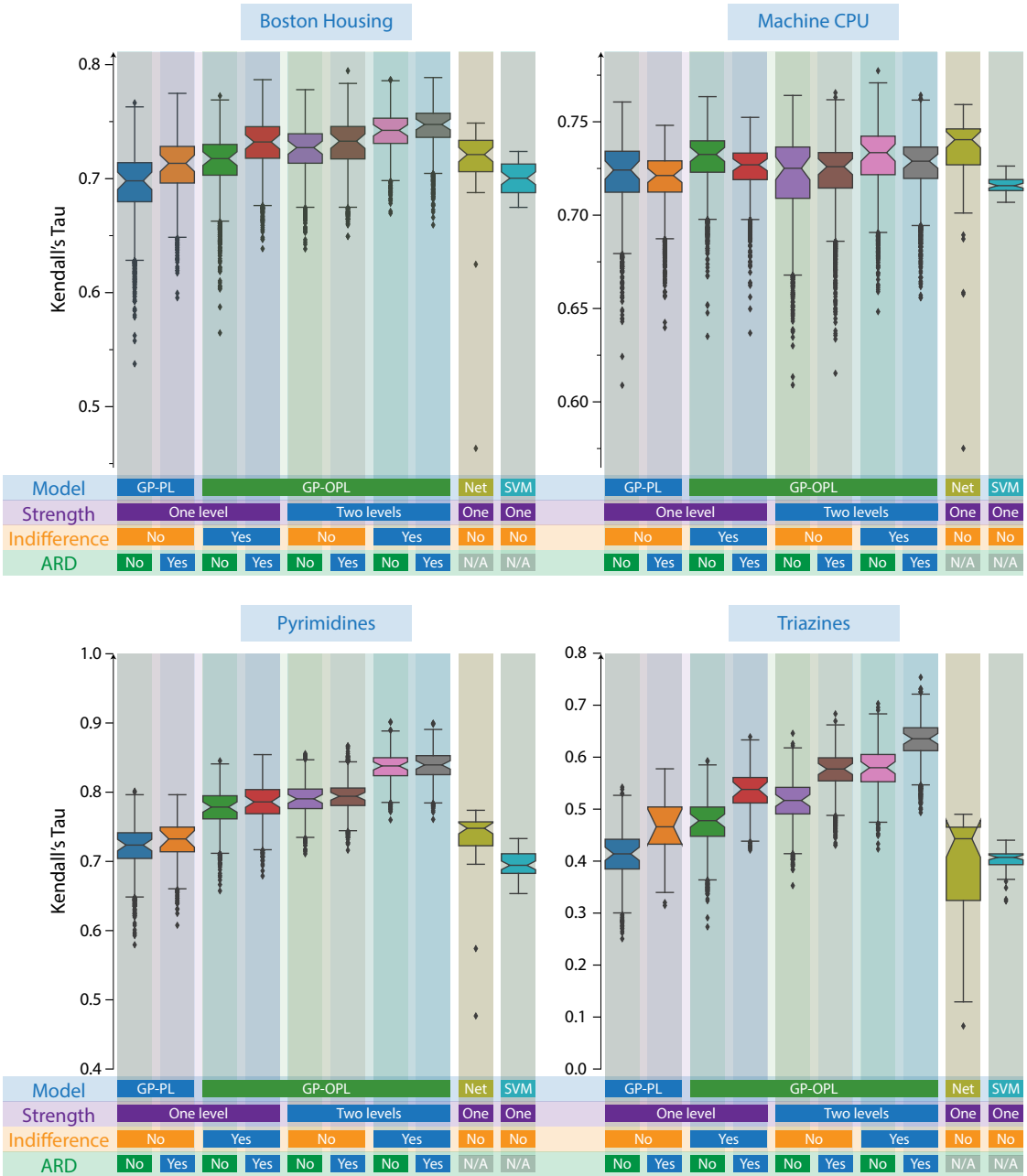


Figure 3.3: Kendall's tau boxplots for Experiment set 1. Kendall's tau range from $[-1, 1]$, where -1, 0, and 1 are perfectly incorrect, random, perfectly correct rank accuracy, respectively. See text for details. For the Pyrimidines and Triazines datasets RankNet had 3 outliers each below the x-axis.

Experiment set 2: Varying strength ratio

For these sets of experiments the ratio between the number of strong and weak pairwise preferences was varied while keeping the total number of observations equal. They are presented in Figures 3.4 and 3.5. As a reference to the two-level strength GP-OPL models both Chu's GP-PL and one-level strength with indifference GP-OPL models were included. All experiments are with an ARD kernel.

AUC Consistently the models without indifference had higher AUC values compared with those models including indifference. Interestingly, the accuracy of two-level strength increased monotonically for all datasets. For the Boston Housing and Machine CPU datasets this was across the whole domain from 1/9 to 8/9; in the Pyrimidines and Trizines datasets the increase was until 7/9 and 6/9, respectively, at which point they decreased.

Kendall's tau The two-level strength model with indifference consistently ranked the latent function f more accurately compared with all one-level strength models and two-level strength model without indifference. Both two-level strength models increased rank accuracy with a maximum between 4/9 and 5/9. Chu's one-level strength without indifference model GP-PL performs worse than all GP-OPL models. Similar to the Kendall's tau results from Experiment set 1, the results for all GP models for the Machine CPU dataset are overlapping with no discernible improvement when two-level strength or indifference are included.

Experiment set 3: Varying training set size

For this set of experiments the number of observations was varied and are presented in Figures 3.6 and 3.7. All experiments are with an ARD kernel.

Both the AUC and Kendall's tau increased at the same rate for all models for the Boston Housing and Machine CPU datasets. In the Pyrimidines dataset the Kendall's tau for all GP models are increasing faster than the comparison models RankNet and RankSVM. This is accentuated in the Triazines dataset whereby there

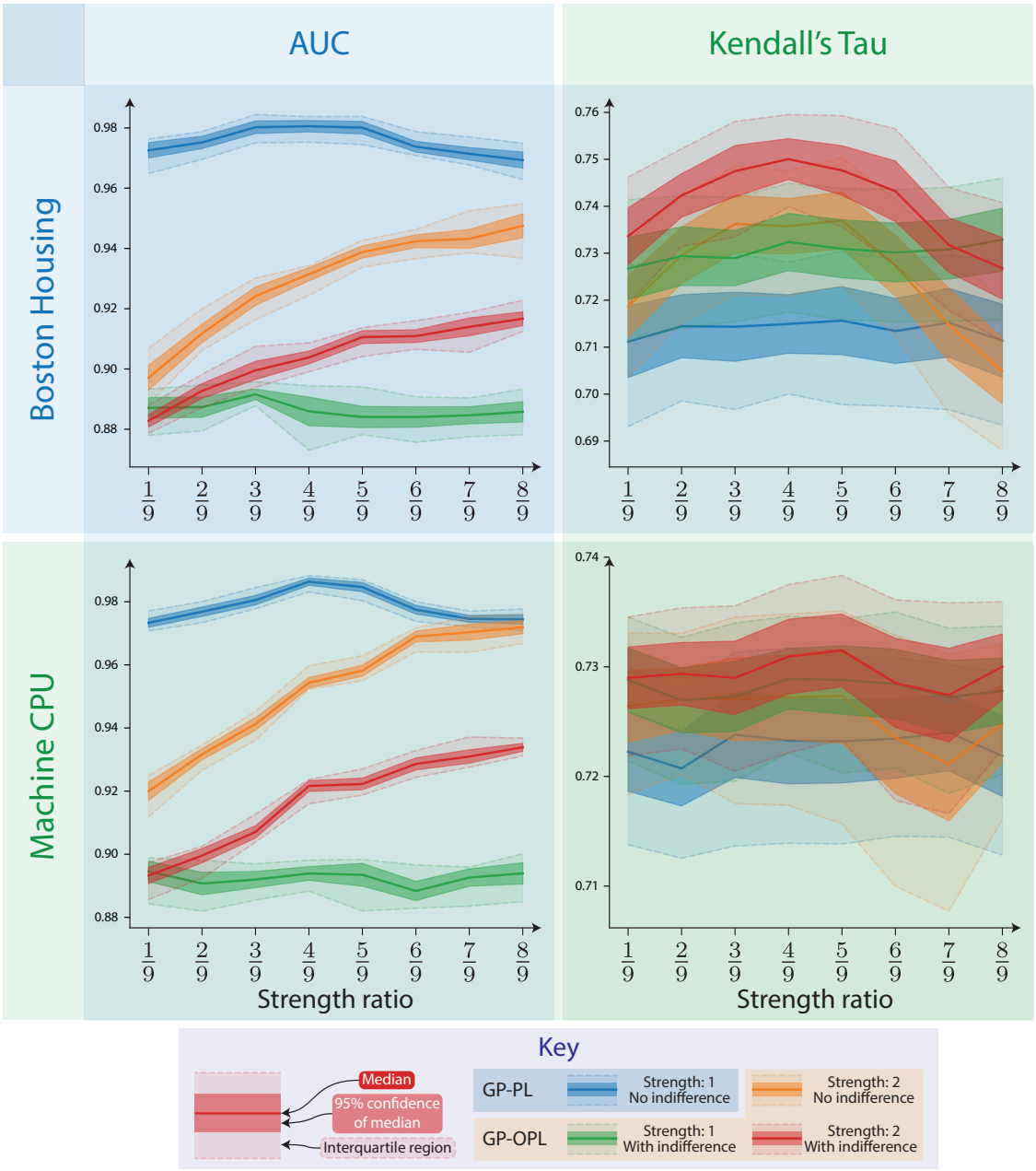


Figure 3.4: Both AUC and Kendall's tau are used to assess the accuracy of GP-OPL compared with GP-PL over a range of different strength ratios. Results for the Boston Housing and Machine CPU datasets are displayed in this figure. See text for more details.

is a linear divergence in the rate of Kendall's tau increase; RankSVM is almost stationary while the rank accuracy for the two-level strength with indifference GP-OPL model continues to increase up to 120 observed training pairs.

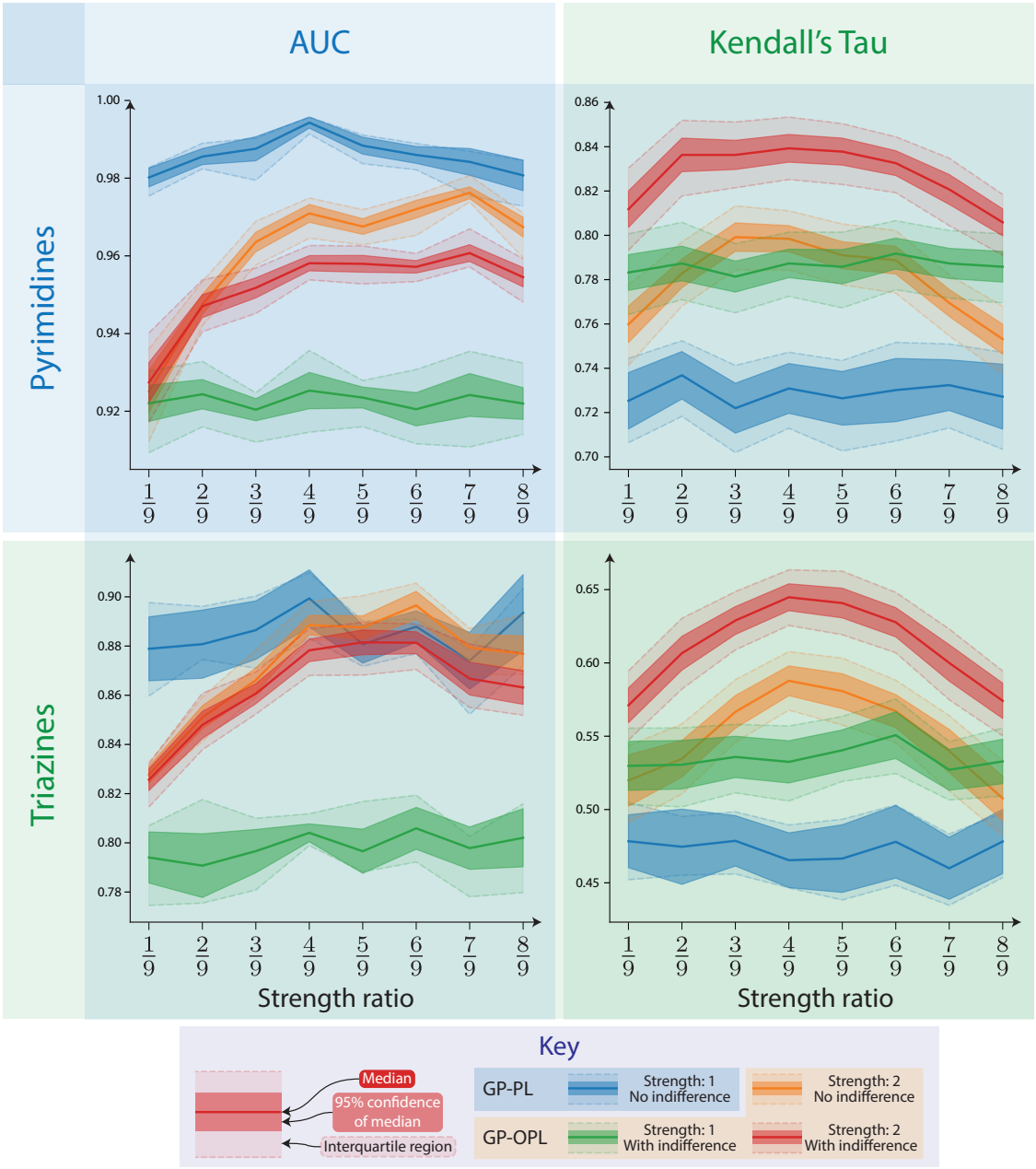


Figure 3.5: Both AUC and Kendall's tau are used to assess the accuracy of GP-OPL compared with GP-PL over a range of different strength ratios. Results for the Pyrimidines and Triazines datasets are displayed in this figure. See text for more details.

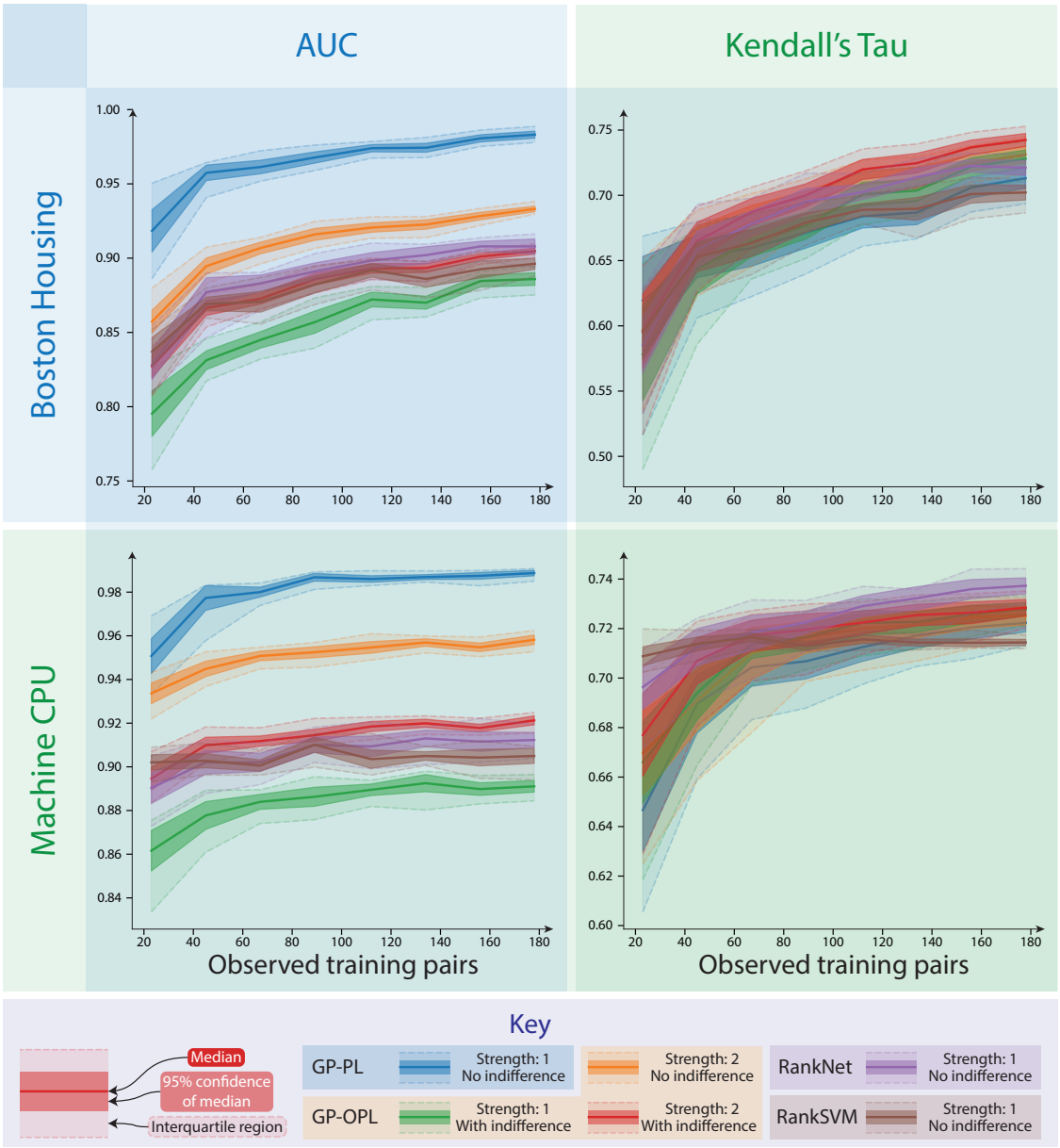


Figure 3.6: Both AUC and Kendall's tau are used to assess the accuracy of GP-OPL compared with GP-PL, RankNet and RankSVM over a range of different training pair set sizes. Results for the Boston Housing and Machine CPU datasets are displayed in this figure. See text for more details.

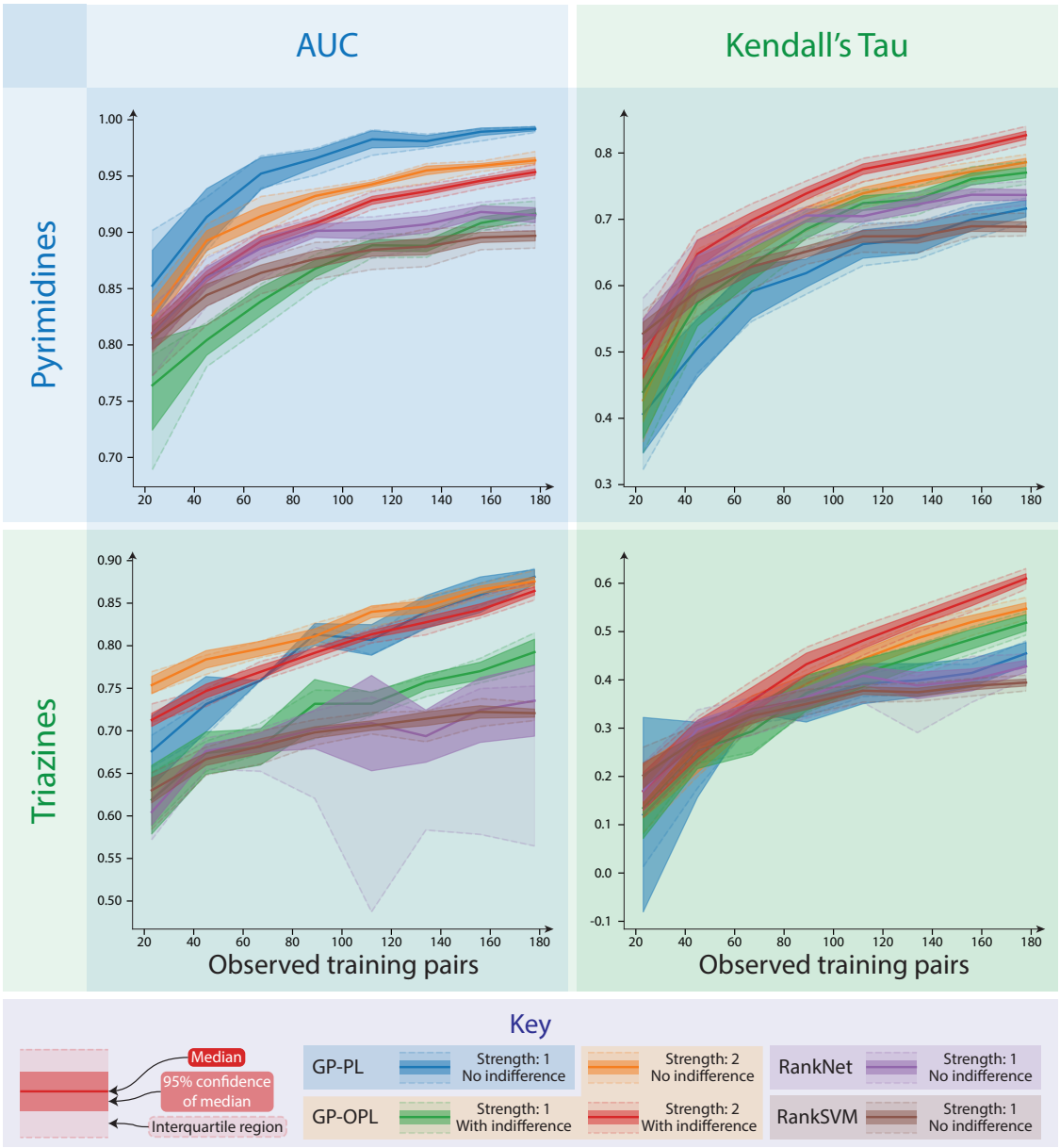


Figure 3.7: Both AUC and Kendall's tau are used to assess the accuracy of GP-OPL compared with GP-PL, RankNet and RankSVM over a range of different training pair set sizes. Results for the Pyrimidines and Triazines datasets are displayed in this figure. See text for more details.

3.3 Heteroscedastic ordinal regression

Having addressed ordinal preference learning we now tackle the complementary problem of including heteroscedastic noise in ordinal regression. This is exemplified when obtaining a self-reported confidence in answer responses to surveys and other human elicited ordinal data. The main question we will be answering in this section is how to include users' self-reported confidence, defined on an ordinal scale, into their self-reported ordinal score. Inclusion of participants confidence is not considered in the homoscedastic ordinal approach. This is important for a more accurate approach.

As detailed at the start of this chapter and in the Background Chapter 2, GP ordinal regression consists of combining ordinal (ordered numerical score) observations with a GP prior. (Chu and Ghahramani 2005b) is currently considered a state-of-the-art approach in this field. However, Chu's model currently does not consider participants' confidence in their responses as mentioned previously. To address this, we developed a model that builds on / differs from existing approaches to GP ordinal regression by incorporating a heteroscedastic noise model. This has two distinct advantages:

1. It increases the ability of participants to express themselves by allowing them to convey varying degrees of certainty about their answers to questions.

2. It enables the model to make predictions both for answers to unasked questions and the participant's predicted confidence for those questions.

3.3.1 Motivation

Let's illustrate the difference between homoscedastic (same) and heteroscedastic (different) observational noise paradigms through this simple example, as shown in Figure 3.8. Imagine a participant is asked a series of questions in a questionnaire and their answers are ordinal. They are also asked to report their confidence of a given answer, which is also on an ordinal scale. Now also imagine that the participant is confident and consistent in their answers in one region of continuous knowledge

space, namely, a low entropy region. Conversely the participant is unconfident and inconsistent in another region, namely, a high entropy region.

If we neglect the participant reported confidence and assume a constant confidence for all regions the model has to balance certainty from one region with uncertainty in another, leading to reduced certainty in predictions.

Including the participants reported confidence and assuming two confidence levels the model more accurately and confidently makes predictions.

It should be noted that if in a different scenario whereby the participant states their certainty but isn't locally consistent we would want the model to discount their confidence labels.

3.3.2 Model description

Here we now specify mathematically the details of our GP-HOR model.

Let the training dataset of n observations be: $\mathcal{D} = \{(\boldsymbol{x}_i, y_s^{(i)}, y_c^{(i)})|i = 1, ..., n\}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_s^{(i)} \in [1, ..., r]$, $y_c^{(i)} \in [1, ..., q]$, d is the dimensionality of the data, r is the total number of ordinal scores, and q is the total number of ordinal confidence levels. For clarity let the training set for scores be $\mathcal{D}_s = \mathcal{D}$ and the training dataset for confidences be $\mathcal{D}_c = \{(\boldsymbol{x}_i, y_c^{(i)})|i = 1, ..., n\}$.

The score and confidence labels are defined as:

$$y_s = \begin{cases} r & a_{r-1} < z_s \leq a_r \\ \vdots & \vdots \\ 1 & a_0 < z_s \leq a_1 \end{cases} \qquad (3.19) \qquad y_c = \begin{cases} q & b_{q-1} < z_c \leq b_q \\ \vdots & \vdots \\ 1 & b_0 < z_c \leq b_1 \end{cases}, \qquad (3.20)$$

where z_s is the latent score value and z_c is the latent confidence value, $[a_0, \ldots, a_r|a_i \in \mathbb{R}]$ are the latent score threshold values and $[b_0, \ldots, b_q|b_i \in \mathbb{R}]$ are the latent confidence threshold values. Initially, in the noise free case observations of score and confidence are independent, such as:

$$z_s = f(\boldsymbol{x}) \qquad (3.21)$$

$$z_c = g(\boldsymbol{x}), \qquad (3.22)$$

where $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are the observational noise free latent functions of the score and confidence, respectively.

We assume the observations are corrupted by additive Gaussian noise, where the confidence observational noise is independent (i.e. homoscedastic), but the score observational noise is dependent on the confidence value y_c (i.e. heteroscedastic). Stating this explicitly:

$$z_s = f(\boldsymbol{x}) + \epsilon_s(y_c) \tag{3.23}$$

$$z_c = g(\boldsymbol{x}) + \epsilon_c. \tag{3.24}$$

More specifically $\epsilon_s(y_c)$ and ϵ_c are zero mean and are described as follows:

$$\epsilon_s(y_c) \sim \mathcal{N}\left(0, \sigma_s^2(y_c)\right) \tag{3.25}$$

$$\epsilon_c \sim \mathcal{N}\left(0, \sigma_c^2\right), \tag{3.26}$$

where σ_c^2 is the variance of the confidence observational noise and the score variance term $\sigma_s^2(y_c)$ is:

$$\sigma_s(y_c)^2 = \begin{cases} \sigma_{s:q}^2 & y_c = q \\ \vdots & \vdots \\ \sigma_{s:1}^2 & y_c = 1 \end{cases}. \tag{3.27}$$

The terms $[\sigma_{s:1}^2, ..., \sigma_{s:q}^2]$ are the q monotonically increasing variances of the score. More specifically an affine transformation of the chosen confidence value: $\sigma_{s:i}^2 = \alpha + \beta(i-1)$ for $i \in \{1, \ldots, q\}$ with hyperparameters $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$.

3.3.3 Gaussian Process Heteroscedastic Ordinal Regression Model

This leads us to incorporating the previous relationships into a Bayesian framework. As in Chu et al. we assume $f(\boldsymbol{x})$ has a GP prior and we also assume there is a GP prior on $g(\boldsymbol{x})$:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m_s(\boldsymbol{x}), k_s(\boldsymbol{x})) \tag{3.28}$$

$$g(\boldsymbol{x}) \sim \mathcal{GP}(m_c(\boldsymbol{x}), k_c(\boldsymbol{x})), \tag{3.29}$$

where $m_s(\boldsymbol{x})$ and $m_c(\boldsymbol{x})$ are the mean functions of the score and confidence latent functions, respectively, $k_s(\boldsymbol{x})$ and $k_c(\boldsymbol{x})$ are the kernel functions of the score and confidence latent functions, respectively.

Likelihood

The likelihood of y_s wrt. z_s for score labels and y_c wrt. z_c for confidence labels are described as follows:

$$p(y_s|z_s) = \begin{cases} 1 & a_{y_s} < z_s \leq a_{y_s+1} \\ 0 & \text{else} \end{cases} \quad (3.30) \quad p(y_c|z_c) = \begin{cases} 1 & b_{y_c} < z_c \leq b_{y_c+1} \\ 0 & \text{else} \end{cases}. \quad (3.31)$$

The likelihood of y_c wrt. g demands marginalising out z_c as follows:

$$p(y_c|g) = \int p(y_c|z_s)p(z_s|g, y_c)\, dz_s \tag{3.32}$$

$$= \Phi\left(\frac{a_{y_c+1} - g}{\sigma_c}\right) - \Phi\left(\frac{a_{y_c} - g}{\sigma_c}\right), \tag{3.33}$$

where $\Phi(u) = \int_{-\infty}^{u} \mathcal{N}(\zeta|0, 1)\, d\zeta$.

Similarly the likelihood of y_s wrt. f requires z_c and y_c to be marginalised out:

$$p(y_s|f) = \sum_i p(y_c^i) \int p(y_s|z_s)p(z_s|f, y_c^i)\, dz_s \tag{3.34}$$

$$= \sum_i p(y_c^i) \left[\Phi\left(\frac{a_{y_s+1} - f}{\sigma_s(y_c^i)}\right) - \Phi\left(\frac{a_{y_s} - f}{\sigma_s(y_c^i)}\right)\right]. \tag{3.35}$$

It should be noted that in the training phase $p(y_c^i)$ is assumed to be a delta function around the observed confidence label. This was done in order to simplify the training process.

Inference

As discussed in section 3.2.3 for GP-OPL inference is not trivial due to the non-conjugacy of the likelihood with the prior, which also applies for the HOR problem.

We again employ the variational Gaussian approximation to infer the posterior distribution, as detailed in (Opper and Archambeau 2008). This assumes multivariate Gaussian distributions $q(\boldsymbol{f})$ and $q(\boldsymbol{g})$ can be used to approximate $p(\boldsymbol{f}|\mathcal{D}_s)$ and $p(\boldsymbol{g}|\mathcal{D}_c)$, respectively, where:

$$q(\mathbf{f}) = \mathcal{N}\left(\mathbf{f}; \mathbf{K}_{s:\mathbf{ff}}\boldsymbol{\alpha}_s, \mathbf{L}_s\mathbf{L}_s^T\right) \tag{3.36}$$

$$q(\mathbf{g}) = \mathcal{N}\left(\mathbf{g}; \mathbf{K}_{c:\mathbf{gg}}\boldsymbol{\alpha}_c, \mathbf{L}_c\mathbf{L}_c^T\right), \tag{3.37}$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]$, $\boldsymbol{g} = [g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_n)]$, $\mathbf{K}_{s:\mathbf{ff}}$ and $\mathbf{K}_{c:\mathbf{gg}}$ are the approximate score and confidence covariance matrices, respectively, with elements $\mathbf{K}_{s:\mathbf{ff}}^{(i,j)} = k_s(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\mathbf{K}_{c:\mathbf{gg}}^{(i,j)} = k_c(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The variational parameters $\boldsymbol{\alpha}_s$ and $\boldsymbol{\alpha}_c$ are the variational mean parameters for the score and confidence approximate posterior, respectively, $\mathbf{L}_s$ and $\mathbf{L}_c$ are square lower triangular matrices and the variational Cholesky parameters for the score and confidence approximate posterior, respectively.

The Python libraries GPflow (Matthews, van der Wilk, Nickson, Keisuke. Fujii, et al. 2016) and Tensorflow (Martin Abadi et al. 2015) were employed again for implementation. Within GPflow the terms $\mathbb{E}_{q(f^i)}\left[\log p(y_s^i \mid f^i)\right]$ and $\mathbb{E}_{q(g^i)}\left[\log p(y_c^i \mid g^i)\right]$ are calculated using Gauss-Hermite quadrature (in all experiments the number of Gauss-Hermite points = 10).

During the inference phase the score and confidence posterior were trained independently due to $p(y_c^i)$ being assumed to be a delta function around the observed confidence label.

Prediction We aim to calculate the distributions of the score and confidence labels (y_s and y_c, respectively) at a new unobserved test instance $\boldsymbol{x}_*$. Firstly, as the score and confidence latent functions f and g are independent we can straightforwardly calculate their distributions given a new unobserved test instance $\boldsymbol{x}_*$:

$$q(f^* \mid \boldsymbol{x}_*, \mathcal{D}_s) = \mathcal{N}\left(f^* \mid \mathbf{K}_{s:\mathbf{f}\star}^T \boldsymbol{\alpha}_s,\, \mathbf{K}_{s:\star\star} - \mathbf{K}_{s:\mathbf{f}\star}^T \Sigma_s^{-1} \mathbf{K}_{s:\mathbf{f}\star}\right) \tag{3.38}$$

$$q(g^* \mid \boldsymbol{x}_*, \mathcal{D}_c) = \mathcal{N}\left(g^* \mid \mathbf{K}_{c:\mathbf{g}\star}^T \boldsymbol{\alpha}_c,\, \mathbf{K}_{c:\star\star} - \mathbf{K}_{c:\mathbf{g}\star}^T \Sigma_c^{-1} \mathbf{K}_{c:\mathbf{g}\star}\right), \tag{3.39}$$

where $\mathbf{K}_{s:\mathbf{f}\star} = [k_s(\boldsymbol{x}_1, \boldsymbol{x}_*), k_s(\boldsymbol{x}_2, \boldsymbol{x}_*), ..., k_s(\boldsymbol{x}_n, \boldsymbol{x}_*)]^T$ and $\mathbf{K}_{s:\star\star} = [k_s(\boldsymbol{x}_*, \boldsymbol{x}_*)]$, $\mathbf{K}_{c:\mathbf{g}\star} = [k_c(\boldsymbol{x}_1, \boldsymbol{x}_*), k_c(\boldsymbol{x}_2, \boldsymbol{x}_*), ..., k_c(\boldsymbol{x}_n, \boldsymbol{x}_*)]^T$ and $\mathbf{K}_{c:\star\star} = [k_c(\boldsymbol{x}_*, \boldsymbol{x}_*)]$, and $\Sigma_s = \mathbf{L}_s \mathbf{L}_s^T$ and $\Sigma_c = \mathbf{L}_c \mathbf{L}_c^T$.

Secondly, from the above results we can now calculate the distribution of the

confidence label y_c^* given a new unobserved test instance $\boldsymbol{x}_*$:

$$q(y_c^*|\boldsymbol{x}_*, \mathcal{D}_c) = \int p(y_c^*|g^*) q\left(g^* | \boldsymbol{x}_*, \mathcal{D}_c\right) \mathrm{d}g^* \tag{3.40}$$

$$= \Phi\left(\frac{a_{y_c^*} - \mu_{g^*}}{\sqrt{\sigma_N^2 + \sigma_{g^*}^2}}\right) - \Phi\left(\frac{a_{y_c^*-1} - \mu_{g^*}}{\sqrt{\sigma_N^2 + \sigma_{g^*}^2}}\right). \tag{3.41}$$

In order to calculate the score label y_s^* given a new unobserved test instance $\boldsymbol{x}_*$ we follow a similar process as with the confidence label but with the added step of marginalising out each possible confidence level y_c^*. The probability distribution of a score label y_s^* given test instance $\boldsymbol{x}_*$ and confidence level y_c^* is:

$$q(y_s^*|y_c^*, \boldsymbol{x}_*, \mathcal{D}_s) = \int p(y_s^*|f_s^*, y_c^*) q(f_s^*|\boldsymbol{x}_*, \mathcal{D}_s) \, \mathrm{d}f^* \tag{3.42}$$

$$= \Phi\left(\frac{a_{y_s^*} - \mu_{f^*}}{\sqrt{\sigma_N^2 + \sigma_{f^*}^2(y_c^*)}}\right) - \Phi\left(\frac{a_{y_s^*-1} - \mu_{f^*}}{\sqrt{\sigma_N^2 + \sigma_{f^*}^2(y_c^*)}}\right). \tag{3.43}$$

Finally this leads us to the final expression where confidence labels y_c^* are marginalised out:

$$q(y_s^*|\boldsymbol{x}_*, \mathcal{D}_s) = \sum_i q(y_s^*|y_c^{*:i}, \boldsymbol{x}_*, \mathcal{D}_s) q(y_c^{*:i}|\boldsymbol{x}_*, \mathcal{D}_c). \tag{3.44}$$

3.3.4 Synthetic experiments

In order to investigate the efficacy of incorporating the heteroscedastic noise model to ordinal regression it is key to compare it to the homoscedastic case. Using the Triazines dataset from the Ordinal Preference section we artificially create a set of ordinal score label distributions and confidence label values, which are split up into train and test sets. We use two metrics to assess the efficacy of the predicted score label distribution, namely, a weighted version of AUC and the Wasserstein distance, both of which are described shortly. For completeness the ordinal AUC metric, as described in sub section 3.2.4, is used to assess the accuracy of the confidence labels.

In all HOR experiments we use an isotropic Gaussian kernel also known as an SE kernel defined as:

$$k_{SE}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa^2 \exp\left(-\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}{2l^2}\right), \tag{3.45}$$

where κ^2 is the signal variance and l is the characteristic lengthscale.

Synthetic heteroscedastic ordinal dataset

The Qualitative Structure Activity Relationships (QSAR) Triazines regression dataset (Dua and Graff 2017) was adapted for the purposes of heteroscedastic ordinal regression, i.e. ordinal labels for the score and corresponding confidence labels. The dataset had the form $(\boldsymbol{x}_i^{True}, v_i^{True}) \in \mathcal{D}^{True}$ with $\boldsymbol{x}_i^{True} \in \mathbb{R}^{60}$ and $v_i^{True} \in \mathbb{R}$. All data items $(\boldsymbol{x}_i^{True})$ were linearly scaled to the unit hypercube $[0,1]^d$. The dataset was split into 50 unique disjoint clusters with cluster centers $\{\boldsymbol{x}_j^{Cluster} | j = 1, \ldots, 50\}$ using the K-Means clustering algorithm provided by the Python package Scikit-learn (Pedregosa et al. 2011).

Score distributions To create an ordinal score distribution $p(y_s^{True}|\mathcal{C})$ we mapped the true data values v^{True} for a given cluster $\mathcal{C}$ into r ordinal bins. The bin thresholds were determined by calculating $r-1$ equally spaced quantiles of the empirical distribution of all true data values v^{True} in the whole dataset $\mathcal{D}^{True}$. More specifically including the extremal quantiles of 0 and 1: Score quantiles $= \{0, \frac{1}{r}, \ldots, \frac{r-1}{r}, 1\}$. Given this empirical distribution of $p(y_s^{True}|\mathcal{C})$ we can sample a score value y_s^{True} for a given cluster $\mathcal{C}$.

Confidence labels We derive a confidence label $y_c|\mathcal{C}$ in two steps. Firstly, by calculating the variance $\sigma^2|\mathcal{C}$ of the empirical score distributions $p(y_s^{True}|\mathcal{C})$ for all clusters. Secondly, similar to the method above, we map the variance $\sigma^2|\mathcal{C}$ of a given cluster $\mathcal{C}$ into q ordinal bins. Again the bin thresholds were determined by calculating $q-1$ equally spaced quantiles of the empirical distribution of all variances $\sigma^2|\mathcal{C}$ for all clusters. More specifically including the extremal quantiles of 0 and 1: Confidence quantiles $= \{0, \frac{1}{q}, \ldots, \frac{q-1}{q}, 1\}$. Note that the confidence label is not a distribution but a single ordinal value for each cluster.

Instances Currently the 50 $\boldsymbol{x}_j^{Cluster}$ data points are only representative of the centre of each cluster not the region which each cluster represents. We assume a

cluster can be represented as an isotropic Gaussian distribution with mean $\boldsymbol{x}_j^{Cluster}$ and standard deviations $\sigma_j^{Cluster}$, where $\sigma_j^{Cluster}$ is defined as:

$$(\sigma_j^{Cluster})^2 = \sum_{i \in \mathcal{C}_j} \left(l_{i,j} - \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} l_{i,j} \right)^2, \tag{3.46}$$

where $l_{i,j} = \left| \boldsymbol{x}_i^{True} - \boldsymbol{x}_j^{Cluster} \right|^2$, $\mathcal{C}_j$ contains the indices of each $\boldsymbol{x}^{True}$ instance in the j^{th} cluster.

From this Gaussian distribution and the ordinal score distribution $p(y_s^{True}|\mathcal{C})$ for each cluster we sample $N = 50$ data points from both distributions for each cluster producing a new dataset $(\boldsymbol{x}, y_s, y_c) \in \mathcal{D}$.

Experiments

The total number of scores r and confidences q were varied in order to assess the relationship between the two, more specifically:

$$r = [3, 4, 5] \tag{3.47}$$

$$q = [1, 2, 4], \tag{3.48}$$

resulting in 9 different score-confidence combinations.

The GP-HOR model was trained on a random subset of 25 clusters and tests were done on the remaining 25. This was repeated 5 times. All experiments were then repeated 5 times, which included the resampling of data points within each cluster.

Metrics

The predictions were tested with two metrics, namely, weighted ROC AUC and the 1-Wasserstein distance. Weighted ROC AUC is used to measure the accuracy of the predicted labels of the test score set. As before, AUC is inherently a metric of binary classification, in order to assess the accuracy of ordinal label predictions we compute the AUC for each label in a *one-vs-all* approach as mentioned in (Waegeman et al. 2008) and take the weighted average:

$$\text{Weighted Ordinal AUC} = \sum_{i=1}^{r} p(y_s^{True}|_i) AUC(\hat{\boldsymbol{y}}_i^{Test}, \hat{\boldsymbol{y}}_i^{Model}), \tag{3.49}$$

where:

$$\hat{\boldsymbol{y}}_i^{Test} = \left[\mathbb{I}\left(y_1^{Test} = i\right), ..., \mathbb{I}\left(y_{m_{Test}}^{Test} = i\right)\right] \tag{3.50}$$

$$\hat{\boldsymbol{y}}_i^{Model} = \left[p\left(y_1^{Model} = i\right), ..., p\left(y_{m_{Test}}^{Model} = i\right)\right], \tag{3.51}$$

and y_j^{True} and y_j^{Model} are the test label and model predicted label, respectively for the j^{th} test instance.

The normalised 1-Wasserstein metric was used to compare the true score distribution $p(y_s^{True})$ with the model posterior score distribution $p(y_s^{Model}|\mathcal{D})$. We use this metric as it takes account of how "close" outcomes may be to each other; this is in contrast to other measures such as the Kullback–Leibler (KL) divergence, which only take account of an outcomes' relative probability (Bellemare et al. 2017). The general p-Wasserstein metric $W_p(P, Q)$ is defined as:

$$W_p(P, Q) = \left(\int_0^1 \left|F_P^{-1}(u) - F_Q^{-1}(u)\right|^p \mathrm{d}u\right)^{1/p}, \tag{3.52}$$

where $1 \leq p < \infty$, and $F_P^{-1}(u)$ and $F_Q^{-1}(u)$ are the inverse cumulative distribution functions of distributions P and Q, respectively.

In the analysis of our experiments we will focus on using the 1-Wasserstein metric, also known as the Earth-Mover metric. In our case the distributions P and Q are the ordinal score distribution $p(y_s^{True})$ and the model posterior score distribution $p(y_s^{Model}|\mathcal{D})$, respectively. Both these distributions are discrete over the domain $[1, r]$, therefore the maximum value the 1-Wasserstein can take is $r - 1$, which is equal to moving a probability mass of 1 from the first bin to the last bin. As the Wasserstein metric is also bounded above 0 we can use these two statements to normalise the Wasserstein metric between 0 and 1 leading to our final expression of the Wasserstein metric used:

$$W(y_s^{True}, y_s^{Model}) = \frac{1}{r-1} \sum_i \left|F^{-1}_{p(y_s^{True})}(u_i) - F^{-1}_{p(y_s^{Model}|\mathcal{D})}(u_i)\right| \tag{3.53}$$

Results

Our results are stated in Figure 3.9. A significant positive correlation between median weighted AUC and the number of score levels r can be observed. A more modest positive correlation existed between median weighted AUC and the number of confidence levels q. This implies that for label accuracy having more score levels is more important than having better knowledge about spread of confidence over the score labels.

In contrast, the Wasserstein metric, which measures the minimum probability mass that needs to be moved from one distribution in order to match a target distribution, tells an inverted story. There is a clear negative correlation between the median Wasserstein metric and the number of confidence levels q. A smaller negative correlation exists between the median Wasserstein metric and the number of score levels r. Therefore, introducing more confidence levels increases the accuracy of the posterior density estimate.

There was no significant correlation between the AUC for confidence labels and the number of confidence levels, which is reassuring as this states that the model accuracy is independent of the number of confidence levels q, over the range of q experimented.

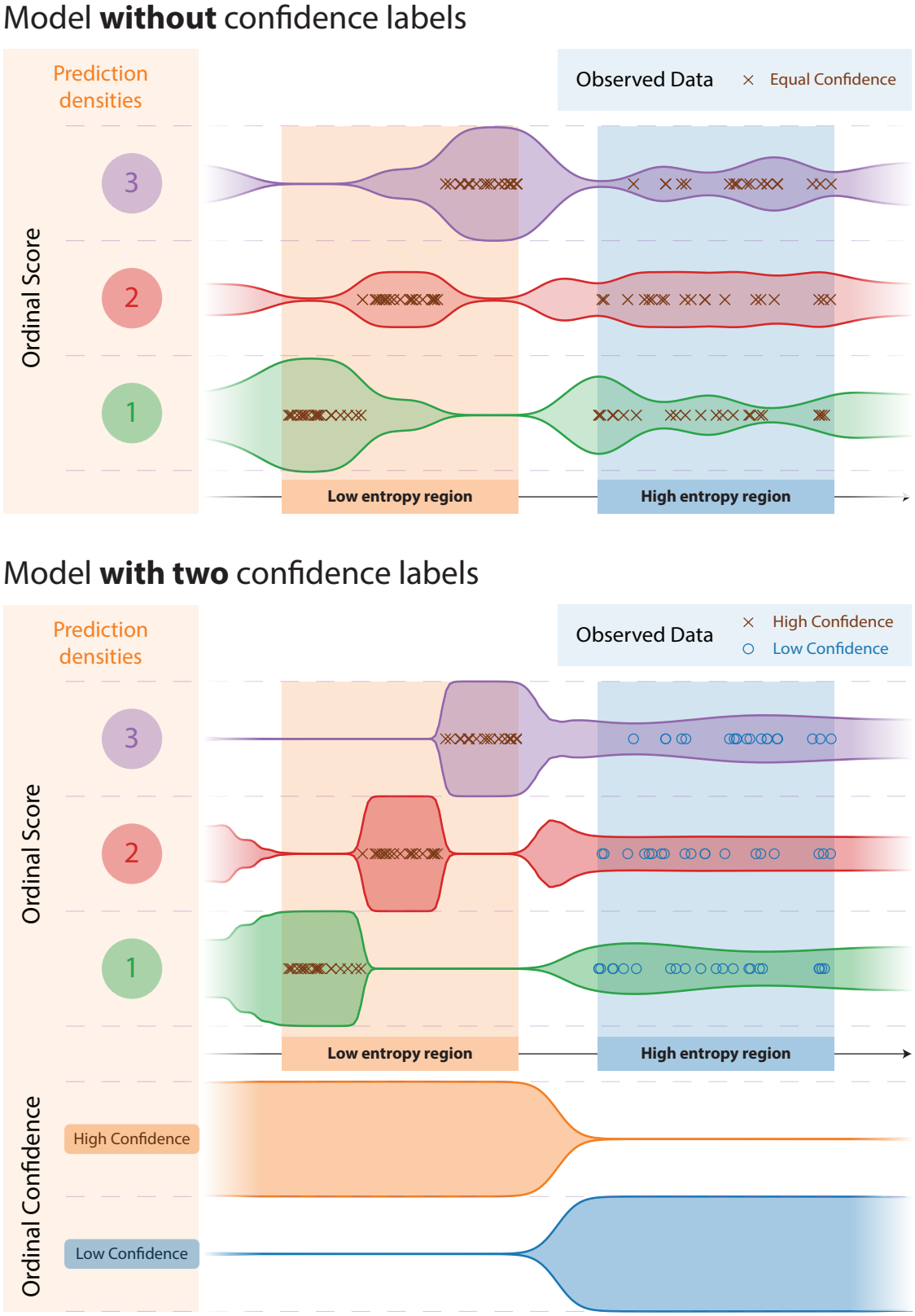


Figure 3.8: One toy dataset applied to two different models, namely, GP-OR and Gaussian Process Heteroscedastic Ordinal Regression (GP-HOR), whereby confidence is provided in the heteroscedastic case. See text for details.

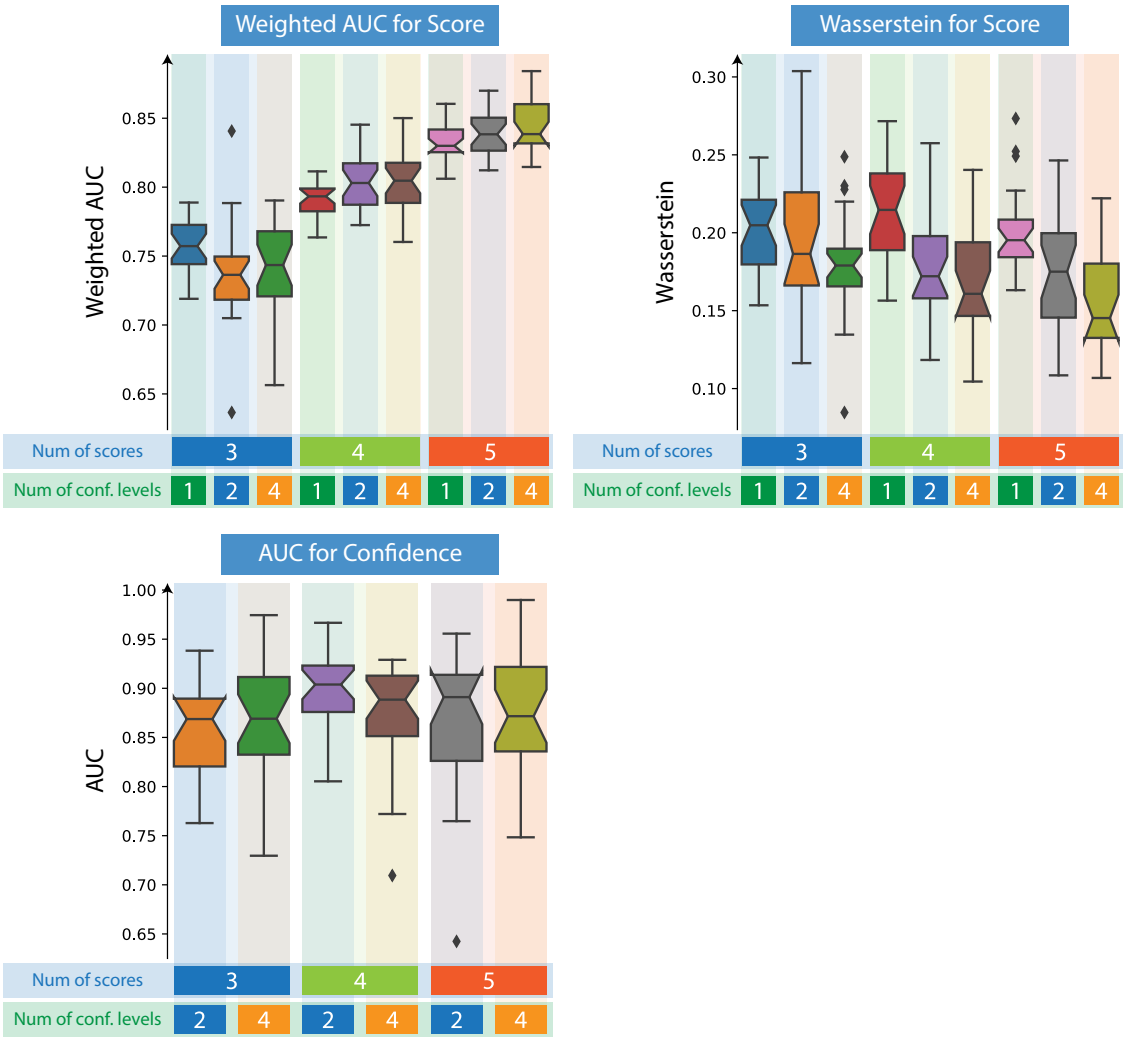


Figure 3.9: Experiment results assessing the efficacy of the GP-HOR model compared with GP-OR. The experiments were carried out on a synthetic dataset derived from the Triazines dataset. Weighted AUC and AUC range from $[0, 1]$, where 0, $1/2$, and 1 are perfectly incorrect, random, perfectly correct ordinal classification, respectively. The Wasserstein metric ranges from $[0, 1]$, where 0 represents two distributions are identical and 1 represents two maximally different distributions.

3.4 Future work

3.4.1 Non-transitive preference learning

A core assumption in both GP-PL and GP-OPL is that there is always a total ordering of items, which enforces transitivity. As noted in (Kahneman 2011) human behaviour is anything but rational and transitive, therefore it would be interesting to explore non-transitivity. That is $A \succ B$ and $B \succ C$ but $C \succ A$, which results in circular reasoning and a function over a single instance doesn't exist that can satisfy those constraints/preferences. A very common example of non-transitivity is in the game: Rock, Paper, Scissors, as shown in figure 3.10.

We can modify the function $g(\boldsymbol{x}_u, \boldsymbol{x}_v)$ from equation 3.1 to a form:

$$g(\boldsymbol{x}_u, \boldsymbol{x}_v) = \tilde{f}(\boldsymbol{x}_u, \boldsymbol{x}_v) - \tilde{f}(\boldsymbol{x}_v, \boldsymbol{x}_u) \tag{3.54}$$

where $\tilde{f}(\boldsymbol{x}_u, \boldsymbol{x}_v) \sim \mathcal{GP}(m(\boldsymbol{x}_u, \boldsymbol{x}_v), k(\boldsymbol{x}_u, \boldsymbol{x}_v))$.

If an ARD kernel is used a measure of how transitive the data is can be obtained. If the importance of the $\boldsymbol{x}_v$ features approaches 0 then $g(\boldsymbol{x}_u, \boldsymbol{x}_v) \to f(\boldsymbol{x}_u) - f(\boldsymbol{x}_v)$, i.e. transitive. Also certain features of $\boldsymbol{x}_v$ may be transitive, therefore along certain hyperplanes the function is transitive, whereas along other hyperplanes the function is non-transitive. Can the underlying manifold that the data lies on be split into transitive and non-transitive sections? So under some transformation the function

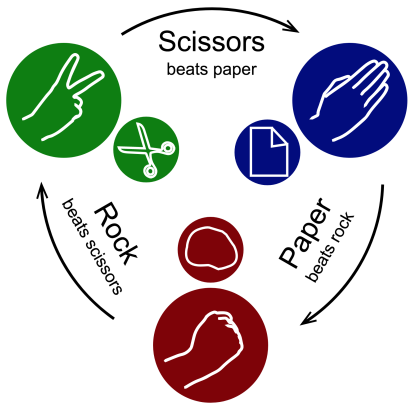


Figure 3.10: Non-transitivity within the game of Rock, Paper, Scissors. (Commons 2019)

can be made transitive? Maybe adding the transitive and non-transitive $g(\boldsymbol{x}_u, \boldsymbol{x}_v)$?

$$g(\boldsymbol{x}_u, \boldsymbol{x}_v) = \{f(\boldsymbol{x}_u) + \tilde{f}(\boldsymbol{x}_u, \boldsymbol{x}_v)\} - \{f(\boldsymbol{x}_v) + \tilde{f}(\boldsymbol{x}_v, \boldsymbol{x}_u)\} \tag{3.55}$$

where $f(\boldsymbol{x}_u)$ is the transitive part and $\tilde{f}(\boldsymbol{x}_u, \boldsymbol{x}_v)$ is the non-transitive part. This leads to an interesting avenue for future research in mixed transitive/intransitive systems. From a global perspective both transitive and intransitive systems can exist together as discussed by (Klimenko 2015).

3.5 Conclusion

This chapter makes the following original contributions:

1. We introduce the novel pairwise preference GP-OPL model which successfully extends GP-PL to include preference strength. This is exemplified by statistically significant improvement of the median Kendall's tau metric relative to GP-PL, RankNet and RankSVM, on a range of synthetically derived datasets.

2. Ordinal heteroscedastic noise in the form of ordinal confidence levels were successfully integrated into the GP-OR resulting in the novel GP-HOR model. Comparing our model to GP-OR experiments consistently showed including confidence levels using GP-HOR increased the posterior density accuracy.

These two models give decision makers a greater understanding of human-centric data in the form of peoples' preferences due to the inclusion of self-reported confidence and ordinal graduation in pairwise relational preferences. Uncertainty of the model and data is captured naturally within our models due to our Bayesian framework. We have shown through experiments that our ordinal preference model is state-of-the-art in performance compared with other published models.

In the next chapter we develop a Gaussian process functional regression model.

4
Functional Regression

Contents

4.1 Introduction

Functional data is a type of human-centric data that extends the description of an instance from a finite feature space to an infinite one. In this chapter we explore and develop a Bayesian model for regressing over functional data, namely, functional

regression. To date, functional regression has primarily been used by the statistics community and been paid little attention by the machine learning community.

As described in the Introduction Chapter 1 interpretability is an important driver of this thesis and we address this in more detail for our functional regression model. For extracting interpretability features in the finite dimensional input case Automatic Relevance Determination (ARD) or shrinkage variables have been used to ascertain the most informative features globally (Guyon and Elisseeff 2003; C.E. Rasmussen and C. K. Williams 2006). More locally we introduced Principled Interpretability for Gradient Evaluation using Bayesian Quadrature (PIGEBaQ) in Chapter 5 to extract interpretable understanding of finite features in a local domain. In this chapter we introduce a Gaussian Process model for continuous additive functional regression that provides interpretable understanding of functional features (i.e. infinite feature) globally.

Normally, in functional regression, functions are taken as the units of observation, and continuous additive functional models are used to map these infinite-dimensional inputs to a scalar response through a non-linear two-dimensional surface. This surface has hitherto been modelled with tensor B-splines, demanding a number of challenging design choices (such as the number of knots) and a difficult learning procedure (e.g. non-differentiability of error function). We propose to instead model the surface with a Gaussian Process, yielding only a single, intuitive design choice (the kernel), and giving principled uncertainty estimates. This Gaussian Process (GP) functional model is demonstrated on synthetic and real-world datasets, and as will be shown in this chapter, provide competitive results.

The main questions we answer in this chapter are:

1. How can uncertainty be incorporated into the functional regression in a principled way?

2. How can we better gain insight into important parts of the functional space with uncertainty?

4.1.1 Overview

Firstly, we will describe functional regression in Section 4.2. We will then go on to describe our GP model in Section 4.3. Synthetic experiments are in Section 4.4, while real-world experiments are in Section 4.5. Future work is in the Section 4.6. Finally the conclusion is in Section 4.7.

4.2 Functional Regression

Functional regression emerges naturally from diverse application fields in science and engineering and is becoming increasingly prevalent (Cardot and Sarda 2005; Ramsay and Silverman 2005; H. G. Müller and Stadtmüller 2005; J. Q. Shi, B. Wang, Murray-Smith, et al. 2007; Fan et al. 2014; Morris 2015; J.-L. Wang et al. 2016). In chemometrics, the functional response is the prediction of a chemical variable on the basis of a digitized signal such as Near Infrared Reflectance spectroscopic information (Aguilera et al. 2013), whilst diffusion tract imaging profiles can be considered functional predictors for performance on an auditory memory test (J. Goldsmith et al. 2011).

The literature predominantly focuses on the Functional Linear Model (J.-L. Wang et al. 2016). Although a number of non-linear models have been proposed over the years, few of these are Bayesian approaches. Of the non-linear approaches, the additive approaches are some of the most powerful (H.-G. Muller et al. 2013; Mathew W McLean et al. 2012). Our work builds on existing additive approaches, enriching them with Gaussian Processes.

Additive models are a popular and powerful statistical modelling tool (Hastie and Tibshirani 1986). They allow one to flexibly model multi-dimensional data through a sum of one dimensional smoothers; ubiquitous examples include logistic regression, linear regression and Generalised Linear Models.

Generalised functional regression emerges as the limit of additive models to infinite dimensional predictors. To see this, consider an additive model over observations of a curve $X(t)$, at increasingly dense time points $t_1, \ldots, t_N$:

$$y = \frac{1}{N}\sum_{i=1}^{N} f(X(t_i), t_i). \tag{4.1}$$

Taking the limit $N \to \infty$ the continuously additive mode (H.-G. Muller et al. 2013), or functional generalised additive model (Mathew W McLean et al. 2012), emerges.

The continuous functional model (H.-G. Muller et al. 2013) is the natural extension of generalised additive models to functional data. Whilst Additive Gaussian Process (Duvenaud et al. 2011) have extended Additive Models, the functional case has not hitherto been addressed.

Functional data analysis is the paradigm where we assume a relationship between a functional predictor, $X(t)$, and a corresponding response y; the units of observations, our inputs, are now the $X(t)$ (Ramsay and Silverman 2005). For an extensive review of functional data analysis, please refer to the reviews (Morris 2015) and (J.-L. Wang et al. 2016). This chapter focusses on function-to-scalar regression.

In classical linear regression, features are assumed to be independent and identically distributed (IID) and the matrix of coefficients $\boldsymbol{\beta}$ is inferred using a pre-chosen loss function. For the situation where the inputs are a function of a latent variable, t, we utilise the functional equivalent:

$$y = \beta_0 + \int_{\mathcal{I}} X(t)\beta(t)\mathrm{d}t + \epsilon, \tag{4.2}$$

where β_0 is a bias term, ϵ is Gaussian noise, $\beta(t)$ is a weighting function; the functional extension of the classical regression coefficient $\boldsymbol{\beta}$, and the interval $\mathcal{I}$ is usually taken as the unit interval $[0, 1]$. The functional regression problem is now framed as solving for $\beta(t)$. Difficulties arise in dealing with the infinite-dimensional object on the right-hand side of Equation 4.2.

This problem is simplified by expressing $X(t)$ and $\beta(t)$ as an orthogonal basis and performing penalised least squares regression. An alternate approach is to perform linear regression on the Functional Principal Components (FPCs) of $X(t)$ (Yao, Hans-Georg Muller, et al. 2004).

A number of non-linear models have been based on the FPC. Full quadratic terms in Equation 4.2 are considered by (Yao and H. G. Müller 2010) whilst the Functional Additive Model (FAM) (H.-G. Müller and Yao 2008) extends functional principal component regression to non-linear models involving additive non-parametric functions of the FPC scores.

Index models are considered by (D. Chen et al. 2011) and non-parametric approaches have been proposed in (Frédéric Ferraty and Vieu 2006; Frédéric Ferraty, Mas, et al. 2006). Bayesian treatments of functional regression have received limited attention, predominantly focusing on the linear case (Crainiceanu and A. J. Goldsmith 2010; James 2002). Function-to-function models have been the focus of previous GP work (J. Q. Shi, B. Wang, Murray-Smith, et al. 2007; J. Q. Shi, B. Wang, Will, et al. 2012; Bo Wang and Jian Qing Shi 2014), whilst (Bo Wang, T. Chen, et al. 2017) use a functional metric as a GP distance measure to directly map the functional predictors to a scalar.

For predictors with high temporal or spatial frequency components, it becomes difficult to capture all the functional dependency in the principal components. As such, we consider the continuous time version of the functional additive. The Functional Generalized Additive Model (FGAM) (Mathew W McLean et al. 2012), and Continuously Additive Model (CAM) (H.-G. Muller et al. 2013), maps the functional predictors through a two dimensional surface:

$$y = \int_{\mathcal{I}} f\{X(t), t\}\mathrm{d}t, \tag{4.3}$$

with $f\{\cdot,\cdot\}$ a smooth bivariate surface, parametrized using tensor B-splines with penalised regularisation (Mathew W McLean et al. 2012) or one-dimensional splines (H.-G. Muller et al. 2013). A Bayesian version with sparse and uncertain functional inputs has also been considered (Mathew W. McLean et al. 2014).

Such an approach requires a multitude of design choices and unnecessary fitting complications: which type of spline functions, how many knots, and a complicated learning procedure. Further, the model provides no uncertainty estimate for predicted values.

We circumvent these difficulties by utilising a GP in Equation 4.3, thus equipping us firstly with a single design choice, the kernel, and allowing us to determine uncertainty bounds around predictions. Further, we gain added model flexibility by the ability to specify a variety of kernels or kernel combinations.

In this chapter, we couple Gaussian Process models with non-linear functional regression, combining probabilistic kernel methods with functional regression approaches. We bring the flexible kernel-based approaches of Gaussian Processes with additive modelling to the problem of functional regression.

4.3 Gaussian Process Functional Model

We introduce the Gaussian Process Functional Generalized Additive Model (GP-FGAM), a new continuous functional additive model, building upon the FGAM introduced previously by (Mathew W McLean et al. 2012) and CAM (H.-G. Muller et al. 2013). The continuous version of the functional additive model is:

$$y = \int_{\mathcal{I}} f\{X(t), t\}\mathrm{d}t + \epsilon, \tag{4.4}$$

where $f\{\cdot,\cdot\}$ is a bivariate non-linear surface, $\epsilon \sim \mathcal{N}\left(0, \sigma_y^2\right)$, and σ_y^2 is the observational variance.

As already highlighted, we focus on the noise free functional predictor case and fill a gap by providing a probabilistic non-linear, time-additive model for functional regression. (Mathew W McLean et al. 2012) assume a separable form for $f\{\cdot,\cdot\}$ using a tensor spline basis, fit using penalised least squares. (H.-G. Muller et al. 2013) similarly use a one-dimensional spline basis. We focus our comparison against the FGAM.

We use a zero-mean GP to model the function $f\{\cdot,\cdot\}$, providing a highly flexible, probabilistic form for the latent function. This also reduces the number of free parameters to determine, leaving us with a single design choice - the kernel. Additionally, we retain the interpretability of the FGAM model, as we are able to infer the latent surface $f\{\cdot,\cdot\}$ as part of the inference process. A further advantage of our model is that we are able to obtain uncertainty estimates for this surface.

These provide the first steps towards answering the first and second questions posed at the start of this chapter.

Following the approach of (Mathew W McLean et al. 2012) we assume that our function factorises in X and t and that therefore the kernel for $f\{\cdot,\cdot\}$ can be written as:

$$\kappa_f((X(t),t),(X(t'),t')) = k_x(X(t),X(t'))k_t(t,t'), \tag{4.5}$$

where $k_x()$ and $k_t()$ are appropriate kernels defined over X and t respectively.

There are other possibilities of how joint kernel κ_f can be structured, such as an additive model (e.g. $\kappa_f((X(t),t)(X(t'),t')) = k_x(X(t),X(t')) + k_t(t,t')$) or neglecting the variable t all together. In the additive kernel case the way X and t interact in an logical OR relationship, e.g. t,t' may be 'far apart'[1] but the two functional evaluations $X(t),X(t')$ may be 'close' leading to a high $k_x(X(t),X(t'))$ and ultimately resulting in $(X(t),t)$ and $(X(t'),t')$being highly correlated.

Conversely, decomposing κ_f into the product of $k_x()$ and $k_t()$, as shown in Equation 4.5, results in a logical AND relations, i.e. $(X(t),t)$ and $(X(t'),t')$ are only highly correlated if both $X(t),X(t')$ *and* t,t' are 'close'.

As the product of kernels maintains the property of locality in X *and* t it was a natural choice for the joint kernel κ_f.

It should be noted that this joint kernel is very similar in structure to the joint kernel employed within a multi-output GP regression model. A multi-output GP model regresses multiple outputs $y_{(l)}$, indexed by l, over their associated feature vector $\boldsymbol{x}_{(l)}$, where the label for the associated output $l = [1,...,n]$, $\boldsymbol{x}_{(l)} \in \mathbb{R}^d$, n is the number of elements in the output vector, and d is the dimensionality of the feature space. In the literature it is common to augment this feature vector to include the output label l leading to $[\boldsymbol{x}_{(l)},l]$ (M. Osborne 2010). Therefore, the kernel function between two augmented feature vectors becomes: $\kappa_y((\boldsymbol{x}_{(l)},l),(\boldsymbol{x}'_{(l')},l'))$. M. Osborne

[1] Given the 2-norm of the difference between variables.

2010 suggests decomposing this joint kernel into the product of two kernels $k_{\boldsymbol{x}}$ and k_l, which results in a kernel very similar to Equation 4.5:

$$\kappa_y((\boldsymbol{x}_{(l)}, l), (\boldsymbol{x}'_{(l')}, l')) = k_{\boldsymbol{x}}(\boldsymbol{x}_{(l)}, \boldsymbol{x}'_{(l')})k_l(l, l'). \tag{4.6}$$

The main difference between the functional kernel presented in this thesis (Eq. 4.5) and the multi-output kernel (Eq. 4.6) is in the type of the input. Each functional model observation contains a vector containing values of t along with a vector of $X(t)$ evaluated at those t values. This is in contrast to the multi-output model input where each observation of a given output l consists of the feature vector $\boldsymbol{x}_l$ and a scalar label l. Therefore, in the multi-output model input there is no information about how one element of $\boldsymbol{x}_l$ relates to another only that the whole vector has the label l. This ultimately results in very different covariance matrix structures.

The next step is to integrate out the variable t from the function f. As f is drawn from a GP with kernel κ_f, the variance for y from Equation 4.4 would be (A. O'Hagan 1992):

$$\mathbb{V}(y) = \int_{\mathcal{I}}\int_{\mathcal{I}} \kappa_f((X(t), t), (X(t'), t'))\mathrm{d}t\mathrm{d}t'. \tag{4.7}$$

Substituting in the specific κ_f from Equation 4.7:

$$\mathbb{V}(y) = \int_{\mathcal{I}}\int_{\mathcal{I}} k_x(X(t), X(t'))k_t(t, t')\mathrm{d}t\mathrm{d}t'. \tag{4.8}$$

The covariance between two input functions $X_i(t)$ and $X_j(t')$ can be readily observed as:

$$\mathbb{C}(y_i, y_j) = \int_{\mathcal{I}}\int_{\mathcal{I}} k_x(X_i(t), X_j(t'))k_t(t, t')\mathrm{d}t\mathrm{d}t'. \tag{4.9}$$

Ideally we would like to compute the integral in Equation 4.9 analytically. Difficulty arises due to the dependence of $k_x(\cdot, \cdot)$ on $X(\cdot)$ which itself depends on t, making such integrations intractable for commonly used kernels.

To circumvent this difficulty we consider observations of $X_i(t)$ to be sufficiently dense; thus without loss of model accuracy we may move away from parametrising the $X_i(t)$ and use the values as direct inputs into $k_x(\cdot, \cdot)$. Computing the covariance in Equation 4.9 is done numerically using a quadrature rule; we utilise Simpson's rule.

4.3.1 Inference and Prediction

Consider being presented with data, $\mathcal{D} = \{(y_1, X_1), (y_2, X_2), \ldots, (y_n, X_n)\}$, with each scalar response value y_i associated with a vector of dense function values $X_i = \Big(X_i(t_1), X_i(t_2), \ldots, X_i(t_d)\Big)$, $X_i(t_j) \in \mathbb{R}$, d is the number of functional evaluations per observation. We concatenate all feature vectors into the feature matrix $\underline{X} \in \mathbb{R}^{n \times d}$, such that the ith row of $\underline{X}$ is X_i. We then define $\mathcal{K}_f(\underline{X}, \underline{X})_{ij} = \int_{\mathcal{I}} \int_{\mathcal{I}} k_x\Big(X_i(t), X_j(t')\Big) k_t(t, t') \mathrm{d}t\mathrm{d}t'$ the full covariance matrix between each of the functional predictors, and $\underline{y} = [y_1, \ldots, y_n]$ the corresponding responses.

Training the GP-FGAM model, which amounts to estimating the kernel parameters, is achieved by maximising the log-likelihood $\log p(\mathcal{D})$:

$$\log p(\mathcal{D}) = -\mathbf{y}^T \left(\mathcal{K}_f(\underline{X}, \underline{X}) + \sigma_y^2 \mathrm{I}\right)^{-1} \mathbf{y}^T \tag{4.10}$$

$$- \frac{1}{2} \log |\mathcal{K}_f(\underline{X}, \underline{X})| - \frac{N}{2} \log(2\pi). \tag{4.11}$$

Given a new observed functional trajectory $X^*(t)$, with its corresponding values $\underline{x}^*$, the prediction is given by the usual GP equations:

$$\mathbb{E}[y^*] = \mathcal{K}_f(\underline{x}^*, \underline{X})(\mathcal{K}_f(\underline{X}, \underline{X}) + \sigma_y^2 \mathrm{I})^{-1} \boldsymbol{y}, \tag{4.12}$$

$$\mathbb{V}[y^*] = \mathcal{K}_f(\underline{x}^*, \underline{x}^*) - \mathcal{K}_f(\underline{x}^*, \underline{X})(\mathcal{K}_f(\underline{X}, \underline{X}) + \sigma_y^2 \mathrm{I})^{-1} \mathcal{K}_f(\underline{X}, \underline{x}^*), \tag{4.13}$$

where $\mathcal{K}_f(\underline{x}^*, \underline{X})$ is a $1 \times n$ matrix with the jth column given by:

$$\mathcal{K}_f(\underline{x}^*, \underline{X})_j = \int_{\mathcal{I}} \int_{\mathcal{I}} k_x(X^*(t), X_j(t')) k_t(t, t') \mathrm{d}t\mathrm{d}t'. \tag{4.14}$$

4.3.2 Latent surface

Use of a GP affords us the ability to construct the underlying surface $f\{\cdot, \cdot\}$ conditioned on observations of y. This is a slightly unusual situation, where we are interested in unobserved function values conditioned on observed integral quantities. Fortunately, by using a GP to represent the surface, we are able to compute the

expected values and obtain an uncertainty estimate for the surface. The joint distribution of $\mathbf{y}$ and $f\{\cdot,\cdot\}$ is:

$$\begin{bmatrix}\mathbf{y}\\ f\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ 0\end{bmatrix}, \begin{bmatrix}\mathcal{K}_f(\underline{X},\underline{X}) + \sigma_y^2\mathrm{I} & \kappa_{yf}(\underline{X},(X^*(t^*),t^*))\\ \kappa_{fy}((X^*(t^*),t^*),\underline{X}) & \kappa_f((x^*,t^*),(x^*,t^*))\end{bmatrix}\right), \tag{4.15}$$

with κ_f as defined in Equation 4.5 and the cross covariance is given as:

$$[\kappa_{fy}((X^*(t^*),t^*),\underline{X})]_j = \int_{\mathcal{I}} k_x\Big(X^*(t^*), X_j(t')\Big)k_t(t^*,t')\mathrm{d}t'. \tag{4.16}$$

The posterior over the surface, f^*, evaluated at (x^*,t^*) is given by:

$$\mathbb{E}[f^*] = \kappa_{fy}((X^*(t^*),t^*),\underline{X})\Big(\mathcal{K}_f(\underline{X},\underline{X}) + \sigma_y^2\mathrm{I}\Big)^{-1}\mathbf{y}, \tag{4.17}$$

$$\begin{aligned}\mathbb{V}[f^*] = {} & \kappa_f((x^*,t^*),(x^*,t^*))\\ & - \kappa_{fy}((X^*(t^*),t^*),\underline{X})(\mathcal{K}_f(\underline{X},\underline{X}) + \sigma_y^2\mathrm{I})^{-1}\kappa_{yf}(\underline{X},(X^*(t^*),t^*)).\end{aligned} \tag{4.18}$$

4.3.3 Numerical Evaluation and Computational Complexity

In order to evaluate the integrals necessary to compute the covariance matrices we utilise a numerical procedure. Although any numerical procedure may be utilised, we find Simpson's rule works well in practice. The essence of Simpson's rule is that quadratic curves are piecewise fitted to every contiguous group of three points and the approximation to the integral is the area under each piecewise quadratic curve. It's important to note that a higher order quadrature rule could have been employed such as Boole's rule which uses a degree 4 polynomial approximation but higher degree polynomials can also result in reduced accuracy. This is known as the Runge phenomenon (Boyd 1992) where the error in the integral grows exponentially for high degree polynomials. Another important feature of Simpson's rule is that it can easily be parallelised, which is needed in order to utilise GPU speedups. :

Generally, applying Simpson's rule, we have:

$$\begin{aligned}
&\int_{t_0}^{t_n} g(t)\mathrm{d}t \\
&\approx \frac{h}{3}(g(t_0) + 4g(t_1) + 2g(t_2) + 4g(t_3) + \cdots + g(t_n)) \qquad (4.19)\\
&= \frac{h}{3}\sum_{j=1}^{n/2} g(t_{2j-2}) + 4g(t_{2j-1}) + g(t_{2j}), \qquad (4.20)
\end{aligned}$$

with $h = (t_n - t_0)/n$ and $t_i = t_0 + ih$. For a function of two variables, $g(t, t')$ we iterate over the integral to obtain Simpson's double quadrature rule:

$$\begin{aligned}
&\int_{t'_0}^{t'_n}\int_{t_0}^{t_n} g(t,t')\mathrm{d}t\mathrm{d}t' \\
&\approx \int_{t'_0}^{t'_n}\left[\frac{h}{3}\sum_{j=1}^{n/2} g(t_{2j-2},t') + 4g(t_{2j-1},t') + g(t_{2j},t')\right]\mathrm{d}t' \qquad (4.21)\\
&= \frac{h}{3}\frac{h'}{3}\left(\sum_{l=1}^{n/2}\left[\sum_{k=1}^{m/2} g(t_{2l-2},t'_{2k-2}) + 4g(t_{2l-2},t'_{2k-1})\right.\right.\\
&+ g(t_{2l-2},t'_{2k}) + 4g(t_{2l-1},t'_{2k-2}) + 16g(t_{2l-1},t'_{2k-1})\\
&+ 4g(t_{2l-1},t'_{2k}) + g(t_{2l},t'_{2k-2})\\
&\left.\left.+ 4g(t_{2l},t'_{2k-1}) + g(t_{2j},t'_{2k})\right]\right), \qquad (4.22)
\end{aligned}$$

where $h = (t_n - t_0)/n$ and $h' = (t'_n - t'_0)/m$.

Letting $g_{ij}(t,t') = k_x\Big(X_i(t), X_j(t')\Big)k_t(t,t')$, where the i,j subscripts refer to the functions being evaluated, we can compute the covariances:

$$\begin{aligned}
&\mathcal{K}_f(\underline{X},\underline{X})_{ij}\\
&= \int_0^1\int_0^1 k_x\Big(X_i(t), X_j(t')\Big)k_t(t,t')\mathrm{d}t\mathrm{d}t' \qquad (4.23)\\
&\approx \frac{h^2}{9}\left(\sum_{l=1}^{d/2}\left[\sum_{k=1}^{d/2} g_{ij}(t_{2l-2},t'_{2k-2}) + 4g_{ij}(t_{2l-2},t'_{2k-1})\right.\right.\\
&+ g_{ij}(t_{2l-2},t'_{2k}) + 4g_{ij}(t_{2l-1},t'_{2k-2}) + 16g_{ij}(t_{2l-1},t'_{2k-1})\\
&+ 4g_{ij}(t_{2l-1},t'_{2k}) + g_{ij}(t_{2l},t'_{2k-2})\\
&\left.\left.+ 4g_{ij}(t_{2l},t'_{2k-1}) + g_{ij}(t_{2j},t'_{2k})\right]\right), \qquad (4.24)
\end{aligned}$$

where $h = \frac{(t_d - t_0)}{d}$, with d the number of time points. Similarly, the cross-covariances are computed as:

$$
\begin{aligned}
&[\kappa_{fy}((X^*(t^*), t^*), \underline{X})]_j \\
&= \int_0^1 k_x\Big(X^*(t^*), X_j(t')\Big) k_t(t^*, t')\mathrm{d}t' \qquad (4.25) \\
&\approx \frac{h}{3} \sum_{j=1}^{d/2} (g_{*j}(t^*, t'_{2j-2})) + 4g_{*j}(t^*, t'_{2j-1}) + g_{*j}(t^*, t'_{2j}). \qquad (4.26)
\end{aligned}
$$

For a function of d points, Simpson's rule scales as $\mathcal{O}(d)$, and Simpson's rule for double quadrature scales as $\mathcal{O}(d^2)$. For n functions, building the $n \times n$ covariance matrix scales as $\mathcal{O}(n^2 d^2)$. Inference in GPs scales as $\mathcal{O}(n^3)$ due to the inversion of the $n \times n$ covariance matrix and prediction as $\mathcal{O}(n^2)$. Therefore the total training complexity scales as $\mathcal{O}(n^3 + n^2 d^2)$, whilst prediction scales as $\mathcal{O}(n^2(1 + d^2))$. Table 4.1 summarises the complexity of the functional GP model compared to its multivariate counterpart.

	Training	Prediction
Functional GP	$\mathcal{O}(n^3 + n^2 d^2)$	$\mathcal{O}\Big(n^2(1 + d^2)\Big)$
GP Regression	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$

Table 4.1: Computational complexity of functional Gaussian Process model compared to the standard Gaussian Process regression. n is the number of training points and d in the number of observations in the functional predictors.

4.3.4 Kernel Choices

The development of GP-FGAM until now has been kernel agnostic. Though we are free to pick any kernel, the purpose of this chapter is to introduce the Gaussian Process functional model and as such we do not pursue the kernel selection problem. In absence of further *a priori* information, we utilise the squared exponential kernel:

$$
k(x, x') = h^2 \exp\left(-\frac{(x - x')^2}{2l}\right). \qquad (4.27)
$$

The hyperparameters are h and l, the output and input length scales respectively.

It should be noted that as the numerical integration method described in Section 4.3.3 samples $X(t)$ at regular intervals of t, there may be pathological results if the choice of kernel isn't carefully considered. E.g. if the kernel k_x has a shorter lengthscale than the distance between two points $X(t)$ and $X'(t')$ that region would have an independent covariance structure which would be incorrect. Also if a periodic kernel is used for k_t it could nullify the locality benefits that were discussed before.

4.4 Synthetic Validation & Calibration

We demonstrate the improved predictive capability of the GP-FGAM on a synthetic example, replicating those in (Mathew W McLean et al. 2012) comparing the GP-FGAM against FGAM; using the implementation of (Mathew W McLean et al. 2012) provided in R. One hundred replicates of data are generated, each consisting of 100 curves sampled at 200 equally-spaced points in the interval $[0, 1]$, with functional predictors defined as:

$$X_i(t) = \sum_{j=1}^{J} \frac{2}{j}[\mathcal{Z}_{1ij}\phi_{1j}(t) + \mathcal{Z}_{2ij}\phi_{2j}(t)], \tag{4.28}$$

with basis functions:

$$\phi_{1j}(t) = \sqrt{2}\cos(\pi jt), \quad \phi_{2j}(t) = \sqrt{2}\sin(\pi jt). \tag{4.29}$$

The coefficients are drawn from a normal distribution $\mathcal{Z}_{hij} \sim \mathcal{N}(0, \frac{4}{j^2})$, $h = 1, 2$. We consider two values of the number of Fourier terms J: $J = 5$ and $J = 500$, with the former resulting in smoother predictor trajectories. Two surfaces are considered for $f\{\cdot, \cdot\}$, one linear, and a non-linear hill surface:

$$f(x, t) = -0.5 + \exp\left(-\frac{x^2}{5^2} - \frac{(t - 0.5)^2}{0.3^2}\right). \tag{4.30}$$

The error variance in the signal for each sample is given as:

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i}^{N}\left(\int_{\mathcal{I}} f(X_i(t), t)\mathrm{d}t - \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{I}} f(X_i(t), t)\mathrm{d}t\right). \tag{4.31}$$

with the resulting Signal to Noise (SNR) defined as σ_y/σ. Four signal to noise values are considered: $\{1, 2, 4, 8\}$.

Each model is trained on 67 curves with 33 used for prediction. We compare the GP-FGAM against the FGAM and a baseline Squared Exponential (SE)–ARD GP model. A SE kernel is used for $k_x(\cdot,\cdot)$ and $k_t(\cdot,\cdot)$. The baseline SE–ARD GP model uses the functional predictor as a long vector of independent inputs.

Performance of each model is measured using the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_{i*} - y_i)^2}. \tag{4.32}$$

Results in Table 4.2 show that the GP-FGAM provides lower RMSE in every case, showing improvement over the FGAM. Additionally the SE–ARD performs poorly, further emphasising the need for functional GP models.

In Figure 4.1 we plot the estimated mean surface for the hill for SNR = 8 alongside the true surface. Visually we recover a good estimate to the true surface, demonstrating the GP-FGAM is able to learn the true surface. This indicates the potential to use the surface as a tool for model interpretability. Edge effects manifest, likely due to the latent surface returning to the zero mean prior.

4.4.1 Probability Calibration

A central claim of the Gaussian Process approach is to obtain better uncertainty estimates of our predicted values. Naturally, better calibrated estimates allows us to make better decisions given our predictions.

Probability plots are generated for each experiment using Z-scores in which we plot the theoretical quantiles of a normal distribution against the ordered observed quantiles of the Z-scores. The Z-scores, are given by $z_{yi} = (y_i - \mathbb{E}[y^*])/\sqrt{(\mathbb{V}[y^*])}$ or $z_{fi} = (f_i - \mathbb{E}[f^*])/\sqrt{(\mathbb{V}[f^*])}$, where y_i and f_i are the true values.

Figure 4.2 shows the probability plot for the scalar response (SNR = 2, J =500, hill surface) and for the latent surface (SNR = 8, J=5, hill surface). A straight line indicates we have correctly calibrated probability estimates for the predicted response, and similar plots are obtained for all experiment settings. We observe

		Models											
		SE–ARD				FGAM				GP-FGAM			
Surface	J	1	2	4	8	1	2	4	8	1	2	4	8
Hill	5	0.177 [0.174 0.181]	0.166 [0.163 0.170]	0.161 [0.158 0.165]	0.158 [0.155 0.162]	0.116 [0.103 0.132]	0.082 [0.076 0.090]	0.060 [0.057 0.064]	0.061 [0.047 0.077]	**0.097** [0.094 0.099]	**0.068** [**0.066** **0.070**]	**0.049** [**0.048** **0.050**]	**0.035** [**0.033** **0.035**]
	500	0.175 [0.172 0.179]	0.170 [0.166 0.173]	0.163 [0.159 0.166]	0.161 [0.158 0.164]	0.101 [0.097 0.107]	0.072 [0.069 0.076]	0.051 [0.049 0.053]	0.048 [0.039 0.063]	**0.091** [**0.088** **0.093**]	**0.064** [**0.062** **0.065**]	**0.044** [**0.043** **0.045**]	**0.032** [**0.031** **0.033**]
Linear	5	1.567 [1.530 1.604]	1.353 [1.320 1.385]	1.240 [1.210 1.270]	1.158 [1.130 1.187]	1.206 [1.175 1.235]	0.923 [0.885 0.972]	0.646 [0.614 0.684]	0.446 [0.432 0.463]	1.172 [1.143 1.200]	**0.850** [**0.827** **0.872**]	**0.582** [**0.567** **0.596**]	**0.403** [**0.394** **0.413**]
	500	1.555 [1.517 1.591]	1.309 [1.278 1.341]	1.208 [1.179 1.237]	1.159 [1.131 1.187]	1.253 [1.211 1.296]	0.875 [0.852 0.898]	0.619 [0.599 0.640]	0.456 [0.429 0.496]	**1.147** [**1.118** **1.175**]	**0.821** [**0.801** **0.841**]	**0.570** [**0.556** **0.585**]	**0.406** [**0.395** **0.416**]

Table 4.2: Synthetic experiments comparing RMSE values of GP-FGAM against FGAM and SE–ARD including the 95% confidence intervals denoted within [.,.]. Significant minimum RMSE values are highlighted in bold. We compare for the hill and linear surfaces, across a range of SNR values (1,2,4,8) and two values of J, corresponding to the number of Fourier components in the functional inputs. The SE–ARD performs poorly, emphasising the need for GP functional methods. The GP-FGAM provides significant lower RMSE values in all but one case, outperforming the FGAM in predictive capability.

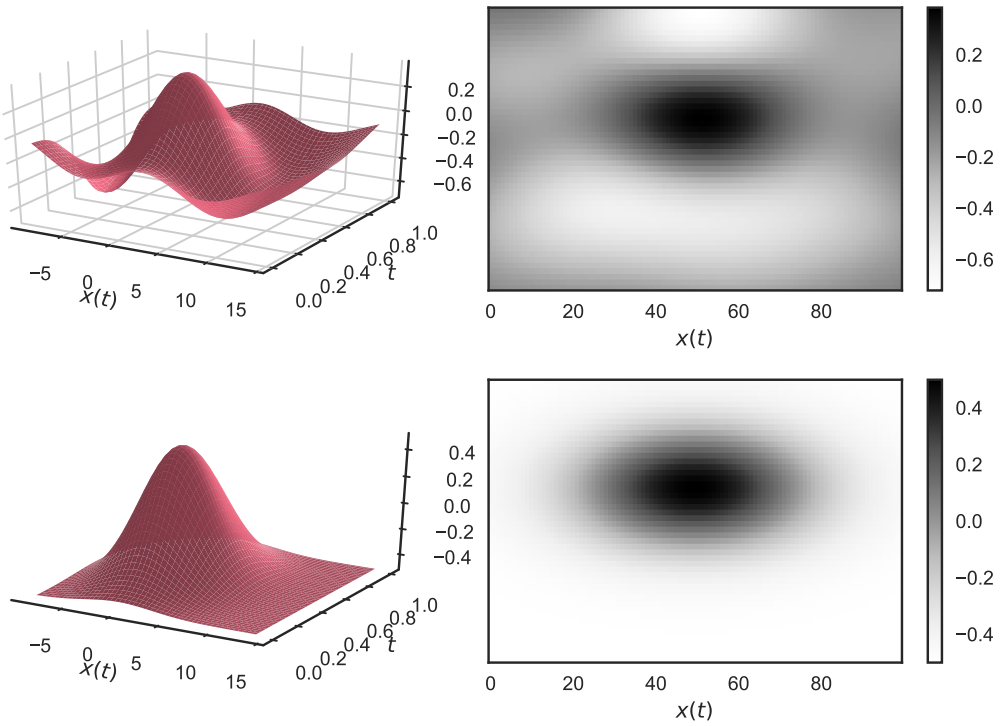


Figure 4.1: Top: the posterior mean surface using GP-FGAM, for SNR=8; we plot the 3D surface and corresponding heat map. Bottom: the true surface. The GP-FGAM is able to recover the correct shape of the latent function. The edge effects are likely due to sparse function observations in those locations and reversion to the zero mean prior.

that we are not correctly recovering the distribution for the latent surface. This could be due to sampling issues and warrants further investigation.

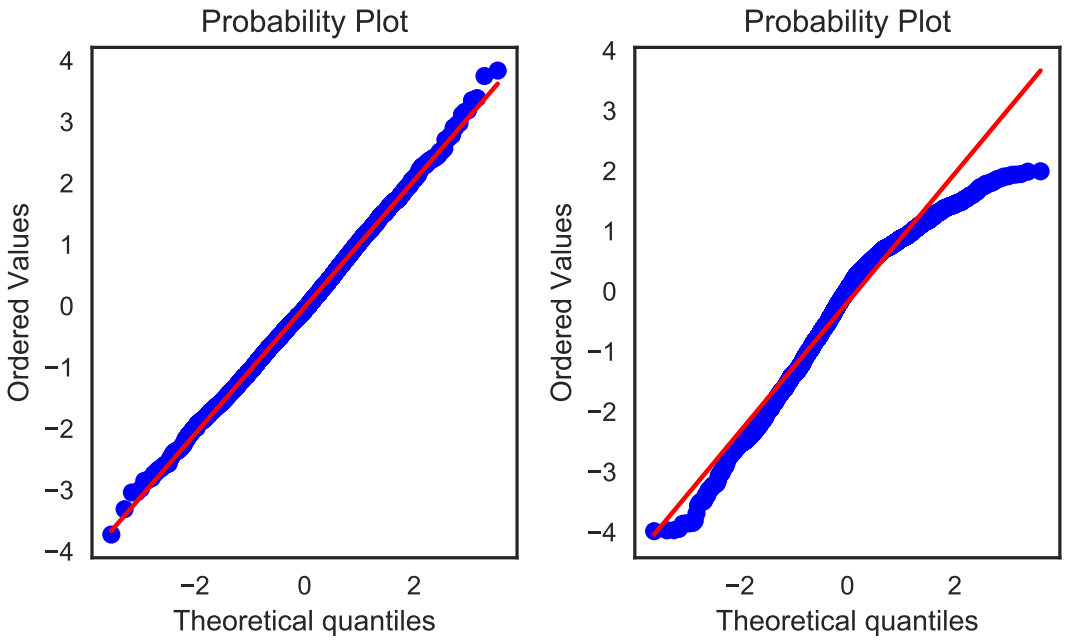


Figure 4.2: Left: probability plot for scalar response : (SNR = 2, J =500, hill surface). A straight line indicates that the predictive variables are normally distributed. Right: probability plot for the surface values: (SNR = 8, J=5, hill surface). We observe deviation from the true quantiles. Similar plots are obtained for other experiment settings.

4.5 Real-World Experiments

In this section we consider a number of challenging real world functional data sets comparing GP-FGAM model against a number of publicly available non-linear functional regression models.

Two linear models are included as baselines: one using the FPC, Functional Linear Functional Principal Component Model (FLM-PCA) (Yao, Hans-Georg Muller, et al. 2004) and one that uses penalised spline regression for the functional basis, Functional Linear Basis Model (FLM-Basis) (Ramsay and Silverman 2005). In addition, we compare against a number of other non-linear and non-parametric models: Functional Quadratic Model (FQM) (Yao and H. G. Müller 2010), FAMs (H.-G. Müller and Yao 2008), Generalised Functional Linear Model (GFLM) (H. G. Müller and Stadtmüller 2005) and Functional Kernel Model (FV) (Frédéric Ferraty and Vieu 2006).

For methods built on Functional Principal Component Analysis (FPCA) (Yao, Hans-Georg Muller, et al. 2004), the number of principal components are selected by choosing the first k that explain 95% of the explained variance. We use the software library PACE[2].

For FV we pick $\min(40, t)$ basis knots for the kernel function: default settings for the available code[3]. FLM-Basis is fit using R, with a smoothing parameter selected by Generalised Cross Validation (GCV) and up to 25 basis functions. For the FGAM we use 6 basis functions for t and 7 for $X(\cdot)$ as in (Mathew W McLean et al. 2012)

Finally, we compare against a baseline GP model: the SE–ARD, treating the functional predictor as a high dimensional input vector with independent features. The SE–ARD GP and GP-FGAM methods are implemented in GPflow (Matthews, van der Wilk, Nickson, Keisuke Fujii, et al. 2016) and we use the SE for all kernels.

We compare each model's predictive capability using a leave one out cross validation exercise on all data sets; we leave one sample out in training, using the held out sample to test predictive performance.

4.5.1 Tecator Data

Tecator (TEC)[4] (Febrero-Bande and Oviedo de la Fuente 2012) consists of data characterising a set of 215 pieces of finely chopped meat. For each piece we observe a single spectroscopic curve which corresponds to the absorbance measured at 100 wavelengths (from 852 to 1050 in steps of 2nm). $X(\cdot)$ is the Near Infra-Red Reflectance (NIR) value at the wavelength. We use the derivative curves of the functional inputs, as they better separate the functional inputs and reduce noise that is present in the NIR measurements (F. Ferraty et al. 2013). The functional inputs are shown in Figure 4.3. The scalar outputs are the measured Fat (FAT), Water (WAT) and Protein (PRO) content, obtained by analytic chemical processing.

The TEC are difficult to compute with GP-FGAM as they require extensive computational resources to optimise the hyperparameters. In order to overcome

[2]Code downloaded from http://www.stat.ucdavis.edu/PACE/
[3]Code available from https://www.math.univ-toulouse.fr/staph/npfda/
[4]This data set can be found at http://lib.stat.cmu.edu/datasets/tecator

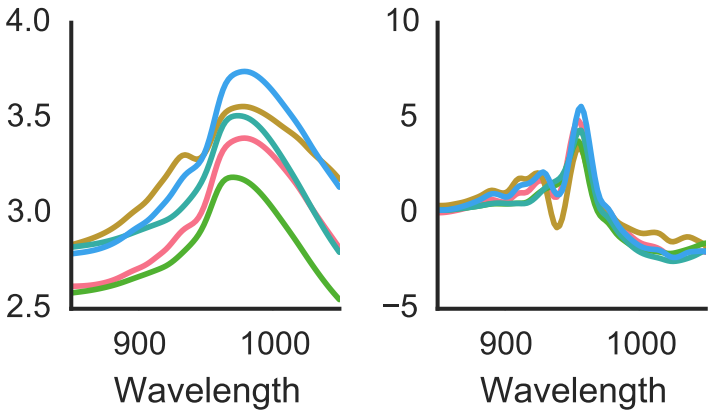


Figure 4.3: TEC functional inputs. Left: TEC, right: derivative. The x-axis is wavelength (nm), the y-axis the spectrometric reading. Each colour is one sample. Clear functional form is observed.

this, we use subsamples of the functional trajectories to train our model and then use those model parameters to predict on the full dataset. We train with 50 observations (out of 100) for the functional predictors and compute the leave-one-out predictions with the full trajectories.

The scalar responses, which are between 0% and 100%, are transformed to real valued outputs using the logit function. Each predictor is shifted to the unit interval $[0, 1]$. The values of the functional predictors and response are whitened: we remove the mean and normalise the values to a standard deviation of one.

4.5.2 Diffusion Tensor Imaging Data

From the John Hopkins University, diffusion tensor imaging was used to collect Fractional Anisotropy (FA) and Mean Diffusivity (MD) tract profiles for a number of patients (J. Goldsmith et al. 2011) and considered in (Mathew W McLean et al. 2012). We have access to the Right Corticospinal (RCST) and Corpus Callosum (CCA) tracts for FA and MD, totalling four data sets, where we aim to regress the Paced Auditory Serial Additional Test (PASAT) score from the functional predictor.

The scalar output is the PASAT score for each patient; a real positive number which we pre-transform through a log transform and then whiten. Figure 4.4 shows data samples from the MD and FA profiles along the CCA and RCST tracts.

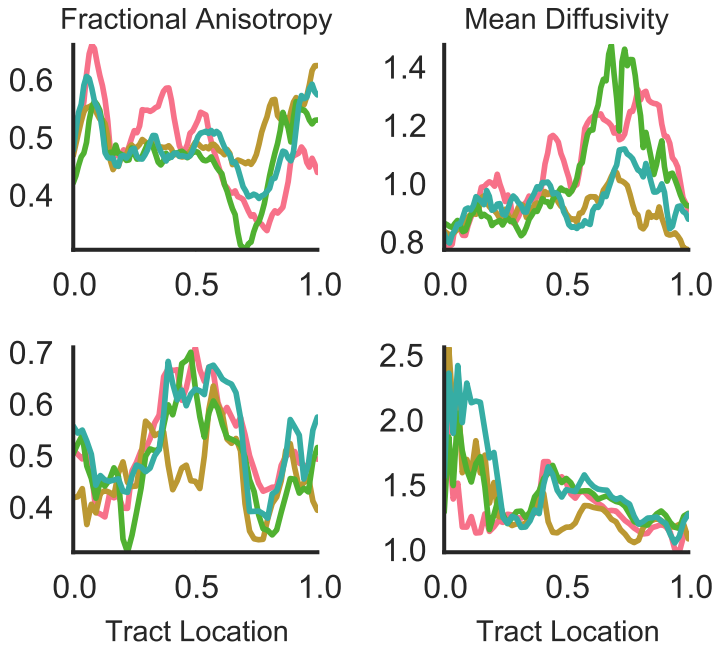


Figure 4.4: Four samples along each tract fo the Diffusion Tensor Imaging (DTI) functional data. Top row: CCA, bottom row: RCST, left image FA, right image MD. We observe distinct functional behaviour along each tract.

4.5.3 Results

The RMSE as shown in Table 4.3, in which smaller values indicate better results. The GP-FGAM gives competitive performance across the board, with low RMSE values, demonstrating good predictive capability. It is particularly notable that it outperforms the FGAM and the ARD in every test. Figure 4.5 plots the absolute surface for the GP-FGAM model for TEC WAT. As expected, the surface mean returns to the zero prior away from the predictors. We note that areas of high value may correspond to salient wavelengths that determine the water content.

	Models								
Datasets	FAM	FLM-PCA	FQM	GFLM	FLM-Basis	FV	FGAM	SE–ARD	GP-FGAM
TEC FAT	53.446 [0.619,92.561]	0.447 [0.374,0.524]	0.260 [0.206,0.316]	0.463 [0.305,0.604]	0.434 [0.360,0.508]	0.249 [0.202,0.296]	0.207 [0.176,0.239]	0.703 [0.631,0.771]	**0.086** **[0.070,0.101]**
TEC PRO	13.064 [0.685,22.603]	0.444 [0.368,0.523]	0.408 [0.340,0.479]	0.510 [0.412,0.606]	**0.207** **[0.183,0.231]**	0.401 [0.325,0.481]	0.398 [0.344,0.449]	0.800 [0.700,0.897]	0.322 [0.273,0.377]
TEC WAT	30.880 [0.619,53.473]	0.241 [0.214,0.269]	0.151 [0.133,0.169]	0.300 [0.202,0.399]	0.222 [0.197,0.247]	0.187 [0.167,0.208]	0.118 [0.100,0.139]	0.777 [0.685,0.866]	0.089 [0.078,0.101]
FA RCST	1.015 [0.815,1.265]	1.110 [0.845,1.392]	2.775 [2.072,3.489]	1.073 [0.852,1.320]	1.044 [0.834,1.293]	1.104 [0.878,1.353]	1.085 [0.846,1.352]	1.512 [1.273,1.748]	1.024 [0.824,1.263]
FA CCA	1.010 [0.856,1.177]	1.019 [0.844,1.206]	1.673 [1.350,1.984]	1.071 [0.911,1.238]	1.027 [0.864,1.211]	0.985 [0.813,1.161]	1.031 [0.856,1.230]	1.130 [0.926,1.336]	0.966 [0.792,1.152]
MD CCA	1.010 [0.851,1.179]	1.006 [0.818,1.204]	1.398 [1.008,1.882]	1.040 [0.872,1.242]	1.037 [0.872,1.210]	1.013 [0.823,1.203]	1.110 [0.943,1.290]	1.087 [0.908,1.284]	1.008 [0.826,1.192]
MD RCST	1.015 [0.813,1.245]	1.044 [0.803,1.322]	1.487 [1.135,1.868]	1.080 [0.816,1.372]	1.033 [0.824,1.280]	1.099 [0.883,1.344]	1.230 [0.902,1.596]	0.963 [0.700,1.260]	1.040 [0.837,1.270]

Table 4.3: RMSE data experiment results including the 95% confidence intervals denoted within [.,.]. Significant minimum RMSE values are highlighted in bold. The GP-FGAM gives competitive results, either best or top three, across the TEC and DTI datasets.

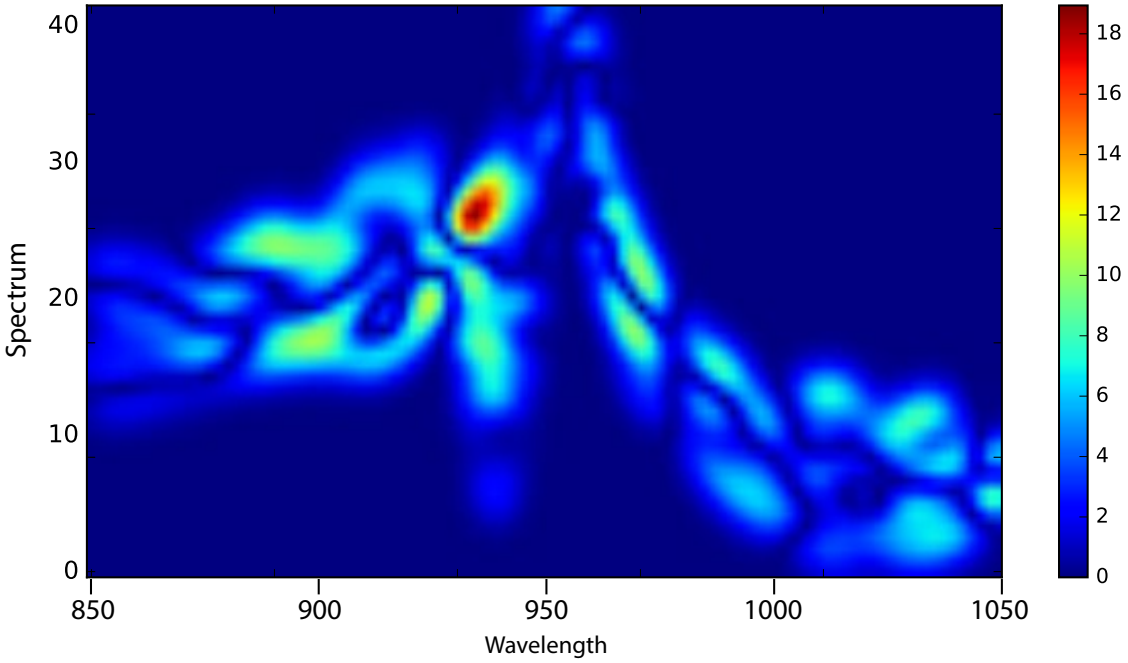


Figure 4.5: GP-FGAM surface plot for TEC WAT. We plot the absolute value of the surface mean, which exemplifies areas of importance for water content prediction.

4.6 Future Work

There are three avenues for future work which could further develop the work outlined in this chapter. Firstly, future work could also involve including noise or sparsity into the functional predictors. Secondly, developing a model for function to function data while naturally incorporating principled uncertainty. Thirdly, there is an interesting possibility for providing a kernel embedding version which could model the underlying functional predictors as GPs and integrate out the t in a tractable way. E.g. Assuming kernel k_x from Equation 4.5 is a linear kernel but the input domain is the predictive distribution from a GP, therefore

$$k_x = X(t)^T X'(t') \tag{4.33}$$

$$= (K_{X^*(t),X} K_{X,X}^{-1} X(T))^T (K_{X^*(t),X'} K_{X',X'}^{-1} X'(T')) \tag{4.34}$$

$$= (K_{X^*(t),X} \alpha_X)^T (K_{X^*(t'),X'} \alpha_{X'(t')}). \tag{4.35}$$

In this instance the requirement for tractability is whether the product of elements of $K_{X^*(t),X}$ can be integrated with respect to t.

4.7 Conclusion

In this chapter we introduced a novel GP model for non-linear functional regression. This probabilistic, non-parametric approach allows us to flexibly model fully observed functional inputs. Therefore, we provided solutions to the two main questions posed at the start of this chapter.

The GP-FGAM surpasses the baselines on synthetic examples using human-centric data in all cases and was able to correctly infer the latent surface as the SNR value increased. The latent surface also gives decision makers valuable interpretable information about the salient areas within the functional predictor. Probability plots indicate that the model provides calibrated uncertainty estimates.

Competitive performance is also obtained on a range of real-world data sets. The utility of the GP-FGAM is demonstrated as it outperforms baseline models in terms of lower RMSE in numerous experiments. The GP-FGAM furthermore outperforms the FGAM, whilst having the additional benefit of well calibrated uncertainty estimates.

In the next chapter we develop a different perspective of human-centric data by extracting human understandable data from complex unintelligible black-box models.

5
Model Interpretability

Contents

5.1 Introduction

In this chapter we focus on the important topic of model interpretability in machine learning. More specifically by extracting human-interpretable meaning from black box models; this is achieved by simplifying the model around local points of interest.

There are three main reasons why this is important: Firstly, there have been a number of high profile instances where an algorithm has been undesirably shown to be unfair, biased and unstable (O'Neil 2016). Secondly, the European Union (EU) and national governments have, or have plans to implement legislation providing legal requirements for decisions made by models to be interpretable (MIRON 2018). Lastly, further insights into a black-box model can inform future research, a proxy for humans instinctive desire to understand the world around us (Vellido et al. 2012). Thus, these reasons drive the research in explainable AI. For example, an insurance company representative who uses a black box algorithm to determine a credit score or to value an asset should be able to explain to a customer and government why they received that score or value.

This has led to an increasing amount of attention being paid to making machine learning models interpretable either internally, as a natural part of the model (e.g. saliency maps (Simonyan et al. 2013)), or externally, by probing the model using other means (e.g Local Interpretable Model-agnostic Explanations (LIME)).

Therefore, providing tools to make machine learning models more interpretable is essential for integration into real-world decision making cases and increase public trust in their use.

Models must be simplified in order to achieve interpretability. As described in the LIME framework (Ribeiro et al. 2016b), interpretability is achieved by approximating a local area of the complex model with a simple human-interpretable model. One of the most simple and interpretable models commonly used is a linear model - a linear relationship between the input and output, i.e. $y \propto \boldsymbol{\beta}^T \boldsymbol{x}$, where $\boldsymbol{\beta}$ is a vector of linear coefficients. The simplicity and interpretability of a linear

model is born from the strong assumptions embedded within it. The first is that the derivative of a linear model results in a constant, in our example that leads to $\frac{dy}{d\boldsymbol{x}} = \boldsymbol{\beta}$, and that all higher order derivatives are zero. So regardless of the value of $\boldsymbol{x}$ when it is perturbed by any amount (e.g. $\delta\mathbf{1}$) the resulting change in y will change by a value only dependent on the constant $\boldsymbol{\beta}$ (e.g. $\delta\boldsymbol{\beta}^T\mathbf{1}$), which is independent of $\boldsymbol{x}$. The potential magnitude of change then depends upon the values of $\boldsymbol{\beta}$, the largest negative/positive elements enable the greatest increase/decrease. Another very important property of linear models is that of element independence, which states that is a change in any element of $\boldsymbol{x}$ has no influence on another. These property of independence greatly simplify understanding by enabling decision-makers to quickly grasp which elements have the greatest or least impact on the output variable without having to think about how they interact or where the change is happening in $\boldsymbol{x}$ space.

Within the LIME Python package a ridge regression model is used to calculate the linear model coefficients. This model doesn't take account of the structure of the data, doesn't report the uncertainty of its assessments, and has no measure of how volatile the surrounding decision boundary is.

That being the case, we have developed a model that agnostically probes a black box model locally to extract a Bayesian probabilistic linear representation. This provides both the average derivative of the target with respect to the feature space, whilst also providing uncertainty measures of how volatile the derivative is locally as well as uncertainty measures of average derivative itself. Extracting these human-centric data we consequently make any black box model with corresponding inputs and outputs understandable, thus further increasing explainability/interpretability and algorithmic fairness.

5.1.1 Interpretability

Definition of interpretability

Interpretability is the degree to which a human can understand the cause of a decision (Miller 2017). In other words, a supervised model can be considered

interpretable if there is a clear relationship, from the perspective of the user, between the input(s) and output(s).

Interpreting a black box model is useful to users as it enables them to understand what changes can be made in order to influence the output of the algorithm. In a complex system with many interacting parts, interpretability also ensures the user is directed to the most significant inputs/mechanisms that can result in changes to the output. Interpretability therefore provides an interface between a user and a model, where the workings of the model are unknown to the user. For example, a user may know that the model performs regression, but this alone would not suffice. For the model to be considered interpretable, the user would need a deeper understanding of the reasons for variation in the output.

If the user has infinite ability to compute and full prior knowledge of a model structure internally then any model can be interpreted exactly. The problem is that a user may only have a restricted ability to compute and possibly does not have full prior knowledge of the model structure. One way around this is to create a surrogate model of the original that is much simpler and one that the user completely understands. This can be done in a global or local sense.

However, due to differences in user knowledge, ensuring the broadest possible interpretability of a model is a difficult task. While universal interpretability is extremely difficult to achieve, working towards this requires us to be mindful of additional questions such as: How can we judge with certainty that a user understands something? Is it possible to say that a user will understand something without checking directly with each user? Can all consequences of using the model be enumerated?

As explained in depth in Section 5.1.1, gradient methods can be used as a tool for interpretability. This is because the change in a parameter of interest (model parameter, outcome value, evaluation metric, or otherwise) can provide useful information for model users, helping them make decisions.

Approaches to the problem of interpretability

We address the problem of recovering principled average gradients by designing a fully-Bayesian pipeline that takes an arbitrary model that (a) outputs a dimension-wise average derivative, (b) can use priors to allow some regions of space to have more or less weighting, (c) provides Bayesian uncertainty of the average gradient, and (d) provides a natural measure of gradient volatility across the space.

Our key insight is that Bayesian quadrature is a natural fit to average gradients in a fully probabilistic way. As a consequence we gain further useful measures such as derivative volatility for gradient interpretability.

We introduce our model: Principled Interpretability for Gradient Evaluation using Bayesian Quadrature (PIGEBaQ) (Principled Interpretability for Gradient Evaluation using Bayesian Quadrature), an end-to-end BQ pipeline for this task. By treating our prior-weighted average gradient as a *random variable*, we recover $p(\int \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x})$, the distribution of our *average* gradient. We use the mean of this distribution as our gradient estimate. In addition, we also use its *uncertainty* and calculate a closed-form expression for its volatility across the space.

The contributions we make in this chapter are to (a) introduce PIGEBaQ and Bayesian quadrature as a natural single tool for this interpretability task, (b) explain how to analytically recover the gradient mean, uncertainty, and volatility, (c) show that it performs comparably to other models used in interpretability literature, and (d) showcase a real-life policy example using PIGEBaQ.

Here, we consider the case of how to make an arbitrary machine learning model interpretable for a real-world policy-maker (politician, CEO, economist, etc.) that wants to understand three questions:

1. **Direction:** Which action should I take in order to have the *most desirable* effect?

2. **Impact:** How *large* is the effect of an action on an outcome?

3. **Confidence:** How *confident* am I about this outcome? (And what would I do if I was not confident?)

In machine learning terms, we care about the *gradient* of a desirable outcome – that is, the vector of derivatives of the outcome with respect to possible policy actions, of which each component is represented as $\frac{\partial\, outcome}{\partial\, action_i}$. For example, an education minister may be interested in how changing actionable variables such as *{student-teacher ratio, enrollment in extracurricular programs, science funding}* relates to the the outcome of graduation rates at individual high schools. Before deciding what policy to enact, the minister would want to know the direction and impact of the actions, and how confident they can be about each action.

However, a fundamental problem in gradient evaluation is that the direction, magnitude, and confidence of the gradient *changes* across the input space. For many non-trivial problems, the common assumption of a linear function from input to output space with uniform variance is a poor answer to a likely nonlinear problem with varying degrees of confidence and regions of space that matter more or less. Dividing the space into regions and approximating the derivative within them requires strong a priori knowledge that is better treated using smooth probabilistic priors.

Overview

The chapter continues as follows: in Section 5.2 an overview of the field so far will be provided. Section 5.3 introduces our model PIGEBaQ and details the mathematics describing our model. Section 5.4 validates and illustrates how PIGEBaQ works. A real-life policy example is presented in Section 5.5. Finally our conclusion and future work are presented in Sections 5.6 and 5.7.

5.2 Related work

We guide the reader by drawing their attention to notable related work which informs and inspires our model. PIGEBaQ is built upon three main ideas: interpretability frameworks, gradient methods, and Bayesian quadrature. Interpretability frameworks guides our thinking on what are the most important questions to be answered. Gradient methods provide a very useful way of delivering interpretability.

Finally, Bayesian quadrature offers a principled framework to bring together gradient methods and interpretability.

5.2.1 Interpretability frameworks

As machine learning techniques become more integrated in the real world, being able to interpret models and their application becomes more important (Vellido et al. 2012). Governing bodies have looked into enforcing the need for interpretability (Goodman and Flaxman 2016). Machine learning researchers are beginning to build interpretability into their models from the beginning (Kim et al. 2015). There is no standard framework for interpretability, though some have been proposed (Lipton 2016; Doshi-Velez and Kim 2017). Generally, these frameworks emphasize understanding concepts such as *trustworthiness, causality, generalizability, meaningfulness, fairness* and *informativeness*. There are commonly-used techniques, from bespoke visualization to post-hoc model simplification (Lakkaraju et al. 2016; Hara and Hayashi 2016), to instance- or model-based approaches relating inputs to parameters or outputs (Kumar et al. 2017; Barratt 2017). A good overview of explaining black box models is provided in (Guidotti et al. 2018).

5.2.2 Gradient methods

Gradient methods are often a desirable interpretability tool, as the change in a parameter of interest (model parameter, outcome value, evaluation metric, or otherwise) can be useful information for decision makers. Analyzing the gradient is not a new method (see (Engelbrecht et al. 1995)) and is sometimes categorized as *sensitivity analysis*, *perturbation methods*, *conditional expectations analysis*, or *influence function analysis* (Koh and Liang 2017; Ribeiro et al. 2016a). Gradient methods for interpretability vary by characteristic. Some are fundamentally *local*, in that they examine gradients within regions of interest, such as in (Ribeiro et al. 2016b; Baehrens et al. 2010; Babibker and Goebel 2017). Other gradient analyses vary by unit (as one might be interested in *relative rank* rather than *magnitude*, or vice versa). Using gradients to interpret machine learning models has been

successfully applied in health outcome prediction (Choi et al. 2016), recommender systems (Hechtlinger 2016), visual decision making (Simonyan et al. 2013), local visual decision analysis (Ribeiro et al. 2016b), and elsewhere. Gradient methods are often powerful for interpretability, yet the difficulties of using a flexible model while retaining interpretable gradients, quantifying uncertainty and volatility, and subsetting in feature space, remain. We address this gap using novel averaging of gradients using Bayesian quadrature.

5.2.3 Bayesian quadrature for principled integration

Bayesian quadrature (A O'Hagan 1991; C. E. Rasmussen and Ghahramani 2003) exploits the flexibility and informativeness of Bayesian priors to perform faster, more accurate, and principled integration. Bayesian quadrature operates by treating an integral as a random variable; by approximating the function to be integrated with a Bayesian prior (like a Gaussian Process (GP) (C. E. Rasmussen and Ghahramani 2003)), the linear operation of integration can be treated probabilistically. Using a model-based approach to inform the sampling procedure can lead to faster and more accurate convergence than standard Monte Carlo methods (Gunter et al. 2014; M. Osborne 2010; M. a. Osborne et al. 2012). For that reason, Bayesian quadrature has been successfully used in sensitivity analysis for electronic circuit design (Pfingsten 2006), decision making in reinforcement learning (Ghavamzadeh et al. 2016), and even predicting storm surges from computationally-intensive hurricane simulations (Toro et al. 2010).

5.3 Model

Here we introduce PIGEBaQ, a model which hitches interpretability onto Bayesian quadrature to provide an end-to-end pipeline from black-box model to interpretable features. We approximate the function of interest with a GP, take the derivative of the function, and perform Bayesian Quadrature on this derivative by using a Gaussian mixture prior.

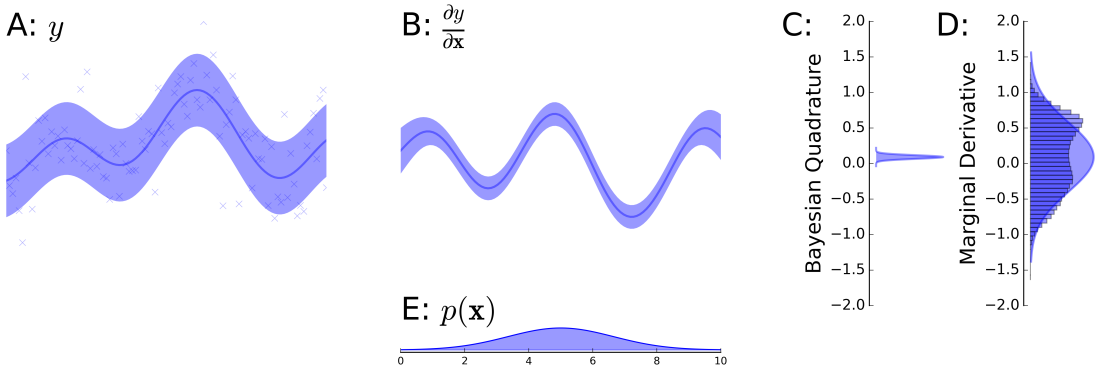


Figure 5.1: The PIGEBaQ pipeline. An underlying model (A) is fit with a GP; (B) we take the derivative of the GP; (C) we find the distribution of $\int \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$, our distribution of the average gradient; (D) we examine the volatility of $\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}$ as a measure of gradient change over the space (the histogram is the empirical distribution and smooth density is our first- and second-moment matched volatility) (E) the prior $p(\boldsymbol{x})$ used in integration.

5.3.1 Derivative quadrature

Let us define f as a function drawn from a GP with mean function $m(\boldsymbol{x})$ and kernel function $K(\boldsymbol{x}, \boldsymbol{x}')$ is written as:

$$f \sim \mathcal{GP}(m(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}')) \tag{5.1}$$

The covariance matrix is generated by a covariance function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, and $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$.

Taking the derivative of f with respect to $\boldsymbol{x}$ gives the derivative function $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$, which is also a GP (C. E. Rasmussen and Ghahramani 2003). We then integrate $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$ weighted by our prior probability estimate of the data:

$$\boldsymbol{I} = \int_{\Omega} \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{5.2}$$

where Ω is the domain of the problem. $\boldsymbol{I}$ is the principled average of the gradient, the quantity we wish to estimate in order to inform decision-making. The covariance between f and $\boldsymbol{I}$ is given by:

$$\mathrm{cov}(f, \boldsymbol{I}) = \int_{\Omega} K_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{x'}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{5.3}$$

where $K_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{x}')$ is the first derivative of the kernel of f with respect to $\boldsymbol{x}$. Similarly the covariance between $\boldsymbol{I}$ and $\boldsymbol{I}'$ is:

$$\mathrm{cov}(\boldsymbol{I}, \boldsymbol{I}') = \iint_{\Omega\times\Omega} K_{\boldsymbol{x}\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}', \tag{5.4}$$

where $K_{\boldsymbol{x}\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{x}')$ is the second derivative of the kernel of f. Explicitly:

$$K_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{\partial K(\boldsymbol{x}, \boldsymbol{x}')}{\partial \boldsymbol{x}}, \quad K_{\boldsymbol{x}\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{x}') = \frac{\partial^2 K(\boldsymbol{x}, \boldsymbol{x}')}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}. \tag{5.5}$$

The attention prior

A mixture of Gaussians is taken as our prior $p(\boldsymbol{x})$,

$$p(\boldsymbol{x}) = \sum_{m=1}^{M} \pi_m p(\boldsymbol{x}|\theta_m)\mathrm{d}\boldsymbol{x}, \tag{5.6}$$

where M is the number of mixtures, m is the mixture index, π_m is the mixture coefficient, and $p(\boldsymbol{x}|\theta_m) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m, \Sigma_m)$ with $\boldsymbol{\mu}_m$, Σ_m as the mean and covariance of mixture m. In practice, any prior density can be used using an importance re-weighting scheme due to PIGEBaQ's probabilistic nature (Gunter et al. 2014). We use a mixture of Gaussians because it is tractable and generally applicable.

The prior is valuable because it allows a policymaker to weight different areas of the space based on how informative they should be to the true gradient. This need arises naturally in many policy cases. In the schools example, some schools in input space have more students and thus policy decisions affect more students. If all areas in the space are equally valuable, one can use a uniform prior. Figure 5.2 captures the intuition for specifying a prior when making policy decisions across groups.

5.3.2 Posterior distribution of the integral

As in previous chapters we are free to pick any kernel, the purpose of this chapter is to introduce our model agnostic PIGEBaQ model and as such we do not pursue the kernel selection problem. In absence of further *a priori* information and for current ease of calculation, we utilise the squared exponential kernel:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{x}')^T \Lambda^{-1} (\boldsymbol{x} - \boldsymbol{x}')}{2}\right), \tag{5.7}$$

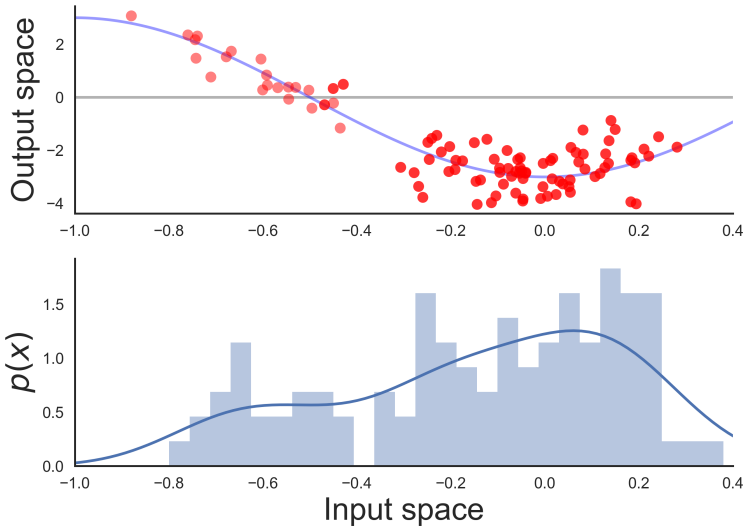


Figure 5.2: Consider some derivative process $\frac{\partial f}{\partial \boldsymbol{x}}$(top) with a positive gradient for groups above the line at zero and negative for those below, with the corresponding distribution of group members (bottom). Suppose its naive average gradient is positive. If we were to use the naive average gradient we would fail to take into account the large negative effect to the larger population for which the gradient is negative. Specifying a prior over the input space allows us to account for differences in the importance of regions in the space. As shown in Section 5.5, there may exist intuitive priors that capture varying importances across input space.

where σ is the output variance hyperparameter and Λ is a diagonal matrix of lengthscale hyperparameters. Derivatives of the kernel are given in Appendix A.

The posterior mean and variance of the average derivative, namely, integral I (Eq. 5.2) conditioned on evaluations of the black-box model f made at $\boldsymbol{x}_i$ is:

$$\mathbb{E}[\boldsymbol{I}] = \sum_{i=1}^{N} \boldsymbol{z}^T \boldsymbol{K}^{-1} f(\boldsymbol{x}_i), \tag{5.8}$$

$$\mathbb{V}[\boldsymbol{I}] = \int\int K_{\boldsymbol{x}\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{x}') p(\boldsymbol{x}) p(\boldsymbol{x}') \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{x}' - \boldsymbol{z}^T \boldsymbol{K}^{-1} \boldsymbol{z}. \tag{5.9}$$

where $\boldsymbol{K}$ is the covariance matrix $\{\boldsymbol{K}\}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, N is the number of black-box model observations, and the term $\int\int K_{\boldsymbol{x}\boldsymbol{x}'}(\boldsymbol{x}, \boldsymbol{x}') p(\boldsymbol{x}) p(\boldsymbol{x}') \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{x}'$ is derived in Appendix A. The components of $\boldsymbol{z}$ are given by:

$$\boldsymbol{z}_n = \sum_{m=1}^{M} \frac{\pi_m \sigma^2 |\Lambda|^{1/2}}{|\Lambda + \Sigma_m|^{\frac{1}{2}}} (\Lambda + \Sigma_m)^{-1} (\boldsymbol{\mu}_m - \boldsymbol{x}_n) \exp\left(-\frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_m)^T (\Lambda + \Sigma_m)^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_m)}{2} \right). \tag{5.10}$$

The second term of the variance is computed using $\boldsymbol{z}$. The first term is a combination of the lengthscale along with the prior means and variances. Note that the variance is independent of the function outputs. Full derivations of both terms are provided in Appendix A.

If we were to have a multi-output black-box model with a vector of output observations $\boldsymbol{y}$ we could employ a co-regionalisation kernel but all outputs would share the same kernel hyperparameters (M. Osborne 2010).

5.3.3 Volatility

Bayesian quadrature allows us to compute a principled average derivative and an uncertainty estimate of the average. However, this uncertainty quantifies only our uncertainty around the estimate of the average derivative, not the variation in the average derivative across the input space. Imagine a situation, such as in Figure 5.3, in which the model is completely certain that the average derivative is equal to 1 over a certain domain. However, across that same domain the derivative could vary drastically, such as it being a sine curve, for example. Thus, we need to consider the volatility of the derivative, which we understand as the variation of the derivative over the input space.

To compute the volatility of the derivative, we marginalise out $\boldsymbol{x}^*$ drawing from the *attention prior* $p(\boldsymbol{x}^*)$. The volatility, denoted as $p(\boldsymbol{y'})$, is calculated as follows:

$$p(\boldsymbol{y'}) = \int_{\mathcal{X}} \mathrm{d}\boldsymbol{x}^* \; p\left(\frac{\partial y}{\partial \boldsymbol{x}^*}\middle| \boldsymbol{X}, y, \boldsymbol{x}^*\right) p\left(\boldsymbol{x}^*\right) \tag{5.11}$$

The volatility term $p(\boldsymbol{y'})$ is intractable as the sum of Gaussian distributions is not a Gaussian or well known distribution, therefore a moment-matched Gaussian approximation is made. This results in the following form of the volatility:

$$p(\boldsymbol{y'}) \approx q(\boldsymbol{y'}) \tag{5.12}$$

$$= \mathcal{N}(\boldsymbol{y'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{5.13}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are derived and presented in Appendix Sections A.3.1 and A.3.2, respectively.

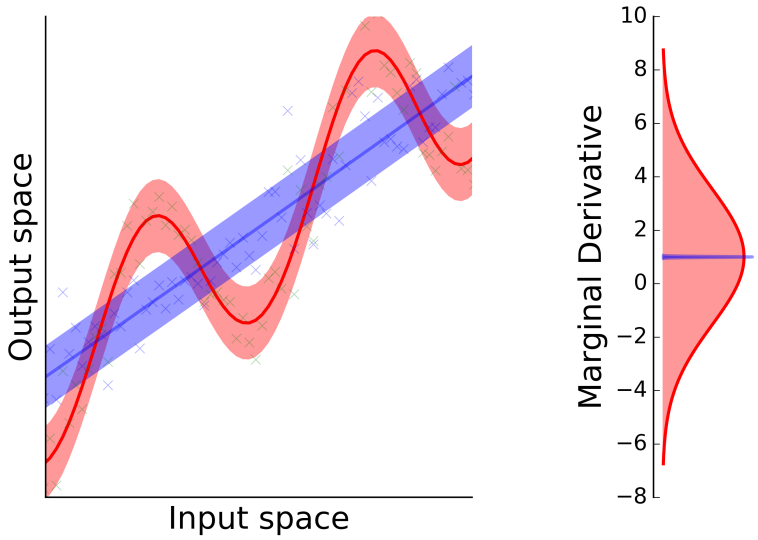


Figure 5.3: An example of when volatility is useful information. The plot on the left shows two underlying functions with the same average gradient. The plot on the right is the distribution of their gradients in the output space. Note that while the gradients are distributed at the same mean, the sinusoidal gradient process is more variable (even negative), implying different policy effects depending on the location of a policy change in the input space.

5.3.4 Further benefits - monotonicity

It is also possible to extend Section 5.3.2 to estimate that the underlying surface is monotonic. Because our integral in Equation 5.2 is fully probabilistic one can use the posterior distribution of the gradient to estimate if the probability is above or below zero. This is important as monotonicity is a desirable quality when deciding on policy actions.

5.4 Validation

Here, we compare PIGEBaQ to a variety of gradient-based interpretation models which are flexible and inflexible. We show that PIGEBaQ and other models comparably recover the true gradient and volatility, and discuss PIGEBaQ's measure of uncertainty and advantages in low-data settings.

5.4.1 Data

We synthesize several functions for which we know the gradient analytically. The functions are specified in Table 5.1. We set hyperparameters $\beta_i \sim \mathcal{U}(0,1)$ and $\gamma \sim \mathcal{U}(0,1)$. We draw n points $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ in d dimensions, where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]$, $\mu_i \sim \mathcal{U}(-1,1)$, and $\boldsymbol{\sigma}^2 = [\sigma_1^2, \ldots, \sigma_d^2]$, $\sigma_i^2 \sim \mathcal{U}(0,1)$.

Table 5.1: Three synthetic functional forms chosen to test PIGEBaQ.

FN.	FORM
f_1	$\boldsymbol{\beta}^T \boldsymbol{x}$
f_2	$\sum_i^M \sin(\pi \beta_i x_i) + \sin(\pi \gamma \prod_i^M x_i)$
f_3	$\sum_i^M \sin(\pi \gamma_i x_i) + \beta_i x_i$

5.4.2 Models

We compare the PIGEBaQ model to four benchmark models. The first is a 3-layer Deep Neural Network (DNN) with layer-wise L_2-regularization. The second model is a standard ordinary least squares linear approximation (Lin). The final model is the LIME Python package (Ribeiro et al. 2016b), which internally uses ridge regression. We selected these three models as they are common in literature when desiring gradient-based interpretability. SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) was considered but was deemed inappropriate as it wouldn't be directly comparative because the result is the Shapley values of each feature not the average gradient. Another point of concern is that SHAP wouldn't be directly comparable to the other methods due to the way SHAP samples around the point of interest $\boldsymbol{x}_*$. All comparisons samples a normal distribution n times around the point of interest $\boldsymbol{x}_*$ and feeds those samples as an argument to the black-box function, then each model processes those data. SHAP on the other hand takes the training examples X (from the black-box model) and the point of interest $\boldsymbol{x}_*$ and creates a synthetic dataset by combinatorially masking the point of interest over each training example. Therefore, SHAP samples the black-box function

Dimension d	Number of samples n
8	256
8	512
16	256
16	512

Table 5.2: Three synthetic functional forms chosen to test PIGEBaQ.

number of training examples $*2^d$ times, where the exponential term explores all combinations of masking the point of interest, which is very sample inefficient and not compatible with our comparisons.

5.4.3 Metrics

We compare models by the Root Mean Square Error (RMSE) of the average gradient relative to the true gradient, and RMSE of the average volatility of the gradient relative to the true volatility.

$$RMSE = \sqrt{\sum_i (y_i^{Model} - y_i^{true})^2}, \tag{5.14}$$

We also compared the full rank order of the average gradient relative to the true gradient using Kendall's tau rank correlation metric, which is described in Section 3.2.4.

5.4.4 Validation experiments

A range of experiments were carried out in order to test and compare the efficacy of the PIGEBaQ to accurately recover the average gradient and average volatility in local regions of space. We explored four combinations of input dimensionality d and number of samples n of each function, as shown in Table 5.2. Each experiment was repeated 20 times.

5.4.5 Results

We present our validation results in Figures 5.4 and 5.5. For all experiments PIGEBaQ performs well, regularly reporting the lowest RMSE and the highest

Kendall's tau for both average gradient and volatility. It should also be noted that as dimensionality increased the degree of improvement of PIGEBaQ over the comparison methods started to diminish consistently. Compared with LIME Python library, a popular method, our method PIGEBaQ uniformly provides better absolute and rank accuracy.

In the most complex function, namely, f_2 the DNN method and PIGEBaQ regularly perform better, although the DNN method has a high range of outcomes especially in low-data scenarios. As expected Lin performed very well on all linear function f_1, this didn't extend well to non-linear functions f_2 and f_3. PIGEBaQ performs roughly in line with the other models recovering the true gradients. The most flexible models (PIGEBaQ and the Neural Network) recover the gradient with varying levels of volatility, while Lin and LIME perform poorly when there is a high degree of complexity.

In the models we've considered, PIGEBaQ is the only flexible model with a fully Bayesian average gradient posterior. Further, its flexibility allows it to recover the true gradient volatility accurately. These two additional pieces of information, as well as accurate average gradients, provide a policymaker with additional and valuable information. The uncertainty and volatility of gradients may significantly change real-world policy choices, as is explored in Section 5.5. In summary, PIGEBaQ is a good balance between low gradient and low volatility error and gives a decision-maker valuable extra information.

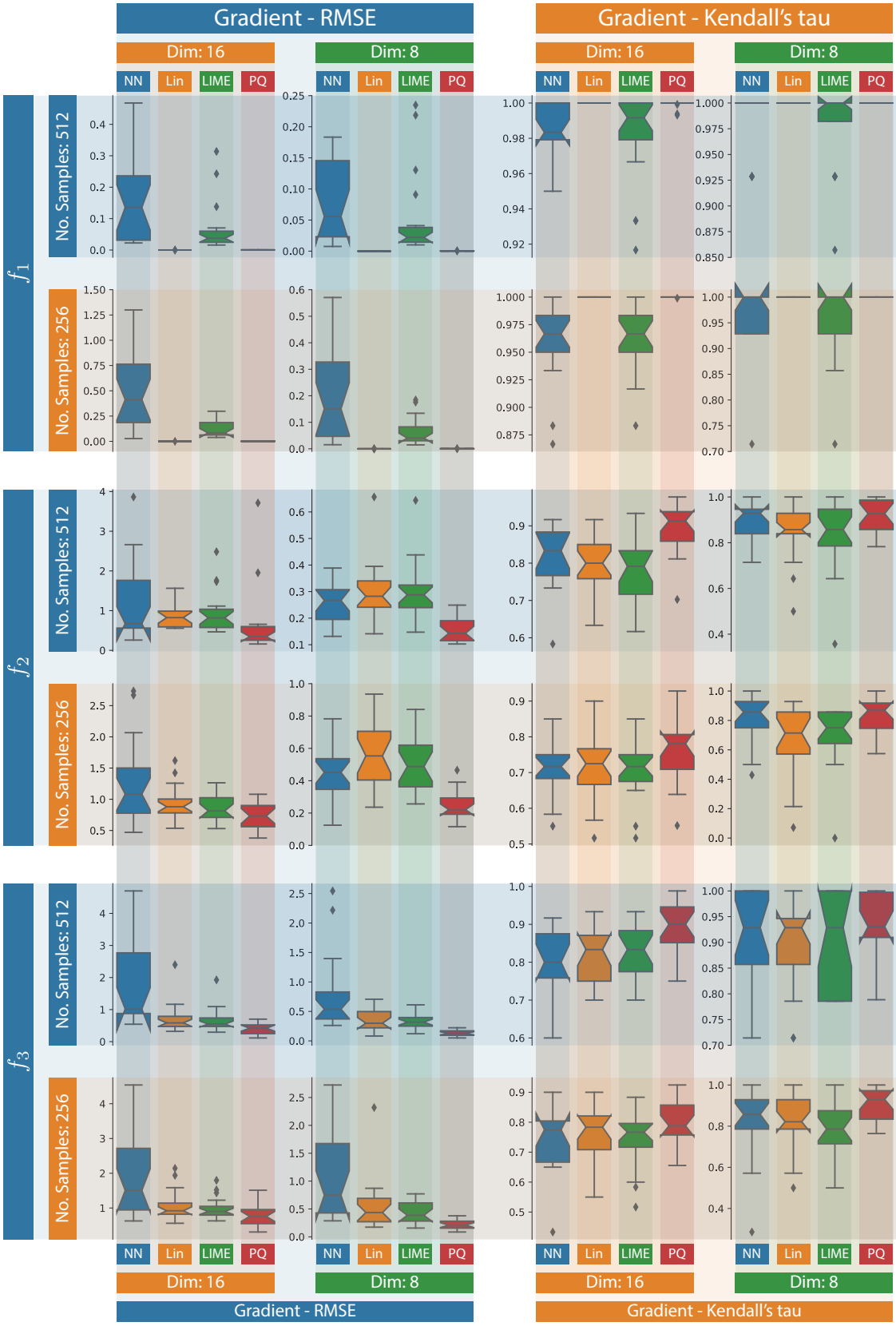


Figure 5.4: Comparison of gradient interpretability methods on synthetic data. Average gradient RMSE and Kendall's tau reported. It should be noted that the acronym **PQ** is a shortened version of PIGEBaQ.

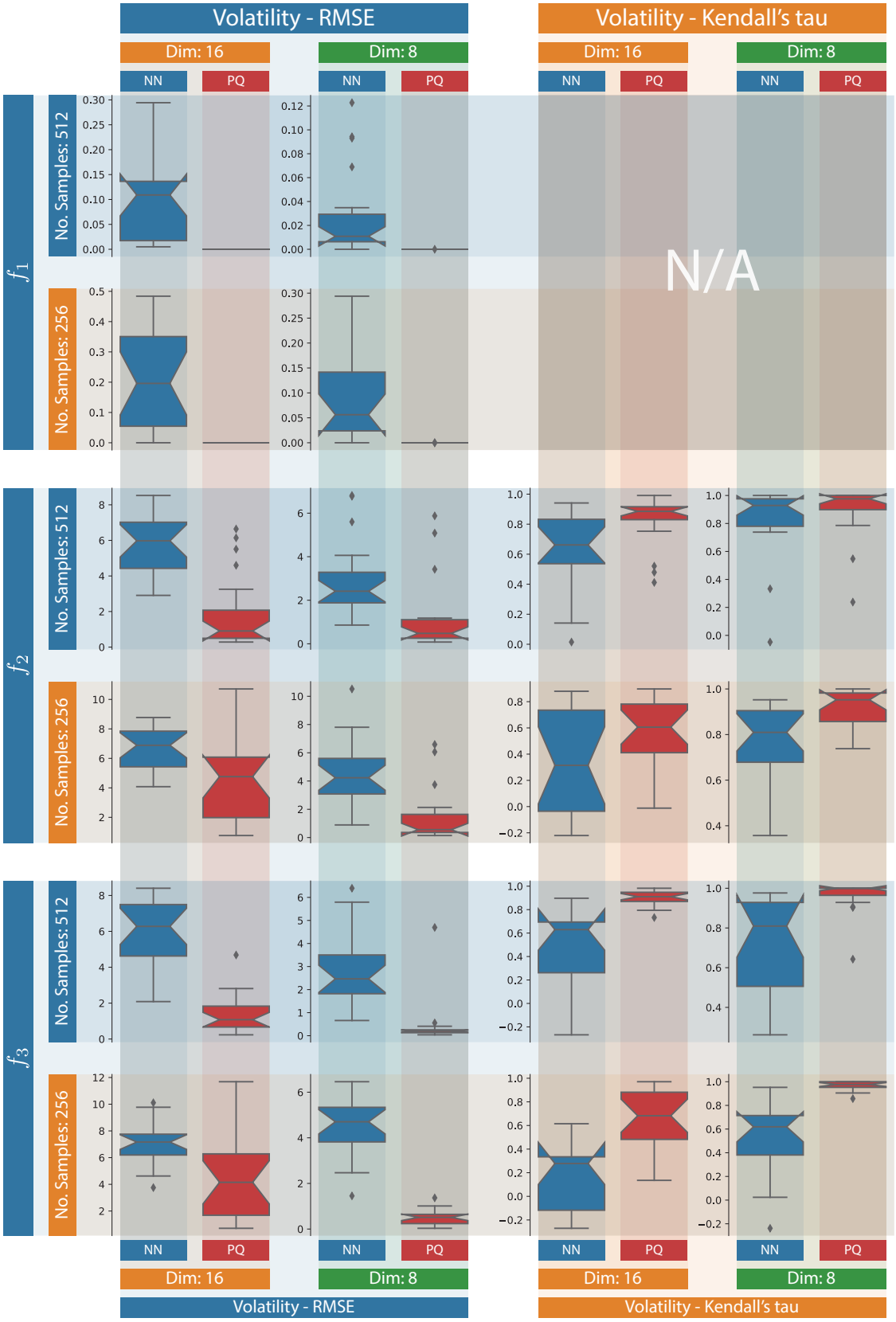


Figure 5.5: Comparison of gradient interpretability methods on synthetic data. Volatility RMSE and Kendall's tau reported. It should be noted that the acronym **PQ** is a shortened version of PIGEBaQ.

5.5 Application

To show PIGEBaQ in practice, and thus demonstrate the value of its features when making a decision, we consider the case of improving post-college economic performance. With almost $1.5 trillion U.S. dollars in student debt held (Fed 2018), and a declining wage premium of college attendance, increasing post-college economic success is an important policy problem. We consider the case of a national education policymaker with the desire to incentivise college-level academic or financial changes to improve economic performance. PIGEBaQ will be used to identify features that would be most useful to help incentivise improved economic performance.

5.5.1 Data

We use data on 4,148 colleges in 2016 provided by the College Scorecard from the U.S. Department of Education (D.O.E. 2017). We use the median student income 6 years after college as our target variable. We use 18 actionable, non-sparse features (having trained with an additional 14 control features for a total of 32 features) that describe the academic and financial makeup of a post-secondary education. Our underlying function f is drawn from a Gaussian Process such that $f : \boldsymbol{X} \to Y$, where $\boldsymbol{X} \in \mathbb{R}^{32}$ is our set of feature descriptors and $Y = y_{income}$ is our economic performance outcome variable.

5.5.2 Prior

In this case, we choose a prior $p(\boldsymbol{x})$ that weights the gradient process by the size of the college within it. This reflects that we would like our average gradient to be more informed by regions in the space which have more students, as the policy goal is to affect the largest number of students. Firstly, we proportionally sample from the input space relative to the number of students in each college. For each college c, we draw $n_c \propto n_c^s$ samples where n_c^s is the number of students in college c. College samples $\boldsymbol{x}_c$ are drawn such that $\boldsymbol{x}_c^{'} \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu_c}$ is college c's feature vector and $\boldsymbol{\Sigma}$ is the diagonal covariance matrix of features over all colleges. We then fit a 10-mixture Gaussian Mixture Model over the sampled space as our prior.

Table 5.3: The action derivatives with the largest magnitudes

	AVG. $\frac{\partial f}{\partial \boldsymbol{x}_i}$
% Degrees in STEM	6.70
% Degrees in Business	6.07
% Degrees in Health	1.84
Instructional expenditures / student	0.38
% Degrees in Law	0.26
Net Tuition/Student	0.19
% Degrees in Arts/Soc Sci	−3.85
% Undergrads on a Loan	−15.54
% Undergrads on Pell Grant	−34.98
% Degrees in Trades	−40.45

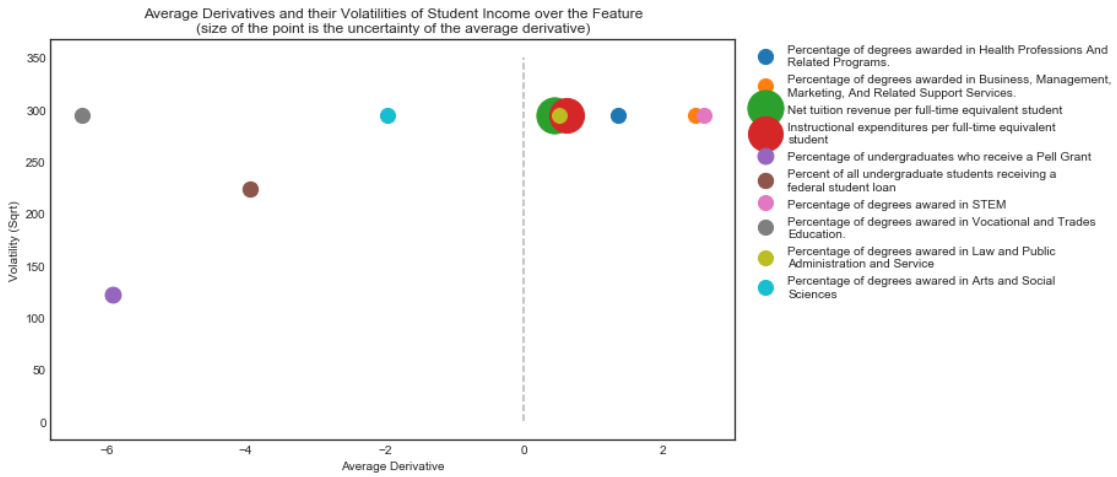


Figure 5.6: Results from running PIGEBaQ on the college data. X is the scaled average derivative, and Y is the gradient's volatility. The size of each point indicates the model's posterior uncertainty about the average derivative.

5.5.3 Evaluating actions

We consider the top dimension-wise average derivatives which are largest in magnitude. Results are shown in Table 5.3.

We see that an understanding of the volatility and the uncertainty allows the policy maker to show that while the action of decreasing *% Degrees in Trades* would have the most positive effect on income, it is highly volatile relative to decreasing the *% of Undergraduates who receive a Pell Grant* but not much

more effective. A policymaker may choose to start there. Alternatively, a policy maker may examine changing *Instructional expenditure per student* rather than the proportion of health-related students, as the model is more certain about the average derivative of the former.

We plot these in Figure 5.6; visualizing as shown may be a useful and intuitive "policy dashboard" for a policymaker.

5.6 Future work

From a technical point-of-view a standard GP with Squared Exponential (SE) kernel will be afflicted by the curse of dimensionality, which has been alluded to in the validation results. Therefore an exploration of other possible kernel functions that may be more appropriate in the high-d scenario could be explored. Also there may be the opportunity to kernalise the SHAP method.

On the application side interesting future work could involve applying these methods to predicting nation state fragility, disease prediction or country-country trade.

5.7 Conclusions

In this chapter the problem of model interpretability is addressed through interpretable gradient averaging. We make the following original contributions:

1. A novel treatment of gradient averaging is presented using Bayesian quadrature resulting in PIGEBaQ. Our method provides a fully principled way to average and interpret average function derivatives.

2. In addition we provide an estimate for the marginal derivative, namely, the volatility of the underlying function. Our results show this to be a very useful tool for assessing movement in the sample.

Our synthetic experiments demonstrate that the quadrature model is able to, with high fidelity, reconstruct the true ranking of features based on their

derivatives. The method was shown to be robust over a range of input dimensions and observed data points.

Using a case study, assessing the probability of economic performance, our model determines the highest average gradient. These features are interpretable and provide compelling evidence that our principled methodology yields results that would help policy makers. The framework we presented provides a methodology for policy makers to understand the three questions of *direction*, *impact* and *confidence* highlighted in the introduction.

This model give decision makers the ability to probe a black-box model thus providing an ability to understand salient features of the problem, consequently helping to inform actions.

In the next chapter we use interpretability concepts from this chapter coupled with the Gaussian Process Heteroscedastic Ordinal Regression (GP-HOR) model of Chapter 6 in the application of Future of Skills.

6
Future of Skills

Contents

6.1 Introduction

Technological advance is causing major changes within the global economy. Labour markets are one place in which these changes are particularly tangible; affecting the daily lives and career trajectories of large numbers of people worldwide. Recently, a growing body of work has been developed that focuses specifically on the automatability of jobs and the resulting changes in employment (for example (Arntz et al. 2016), or (Frey and M. A. Osborne 2017)). While this research has gained substantial attention, there has been a lack of quantitative study of the relationship between skills and demand for skills. The work outlined in this chapter addresses this gap in quantitative research, making a novel contribution to the field by employing a principled Bayesian Gaussian Process Heteroscedastic Ordinal Regression (GP-HOR) model to infer the relationship between skills and demand for skills.

This research has already proven to be highly impactful, and relevant to a variety of influential stakeholders including policy makers, businesses and educators (The Bookseller 2017; Tes 2017; Civil Service World 2017; Public Technology 2017; Education Technology 2017).

The work outlined in this chapter was produced for 'The Future of Skills: Employment in 2030' report, published in September 2017. The report was commissioned by Pearson Plc. and research was conducted in collaboration with Nesta and co-authors Michael A. Osborne, Hasan Bakhshi, Philippe Schneider, Logan Graham and Justin Bewsher. This chapter is designed to illustrate the application of my research to a pressing policy concern, in which the understanding of human-centric data and uncertainty are key. The work presented therefore focuses on the machine learning aspects of the report, where the bulk of my contribution was concentrated. The chapter also focuses on the US case study, however full results for both the US and the UK can be found in the full report (Bakhshi, Jonathan M. Downing, et al. 2017).

6.1.1 Background

Recent debates about the future of jobs have mainly focused on whether or not they are at risk of automation (Arntz et al. 2016; Frey and M. A. Osborne 2017; McKinsey Global Institute 2017; PwC 2017). Studies have also generally minimised the potential effects of automation on job creation, and have tended to ignore other relevant trends, including globalisation, population ageing, urbanisation, and the rise of the green economy.

Our research introduces a novel mixed-methods approach to prediction that combines expert human judgement with machine learning, allowing us to understand more complex dependencies between job features than previously possible. We exploit this enhanced capability to assess complementarities between skills and draw out the implications for new occupations. In addition, our analysis is grounded in an explicit consideration of the diverse and interacting sources of structural change – non-technological as well as technological – all of which are expected to have major impacts on future skills needs. Our identification of the bundles of skills, abilities and knowledge areas that are most likely to be important in the future, as well as the skills investments that will have the greatest impact on occupational demand, provides information that educators, businesses and governments can use for strategic and policy-making purposes.

Finally, unlike other recent studies, our method also makes it possible to quantitatively predict what kinds of new jobs may come into existence. The study challenges the false alarmism that contributes to a culture of risk aversion and holds back technology adoption, innovation, and growth; this matters particularly to countries like the US, which already face structural productivity problems (Atkinson and Wu 2017; Shiller 2017).

6.1.2 Overview

The approach taken in this research is set out in Section 6.2. This introduces the trends taken account of in our research, followed by a description of the foresight workshops, machine learning model and general research design. In Section 6.3,

the data used in our machine learning model is detailed. The methodology used in processing our data is described in Section 6.4. We describe how the Bayesian GP-HOR model developed in Section 3.3 is tailored to the Future of skills application and how all results are derived. Results for the US are presented in Section 6.5. An important aspect is an honest discussion of the limitations of our research along with future work, which is presented in Section 6.6. Concluding remarks are in Section 6.7.

6.2 Approach

For this research we used a three-part approach: trends analysis, foresight workshops, and machine learning analysis. Figure 6.1 illustrates the flow of information in this research from the trends analysis through to the machine learning analysis.

6.2.1 Trends analysis

This section outlines seven key trends we identified as drivers of change that will shape the future of US (& UK) employment. This trends analysis was primarily developed by co-authors (Hasan Bakhshi, Philippe Schneider) and is discussed in more detail in pages 25-28 of the full report (Bakhshi, Jonathan M. Downing, et al. 2017). These broad contextual factors were considered by all co-authors in the process of developing our approach to the research and technical model.

1. **Technological change** This trend takes into account the growth of connectivity, computing power, device ownership and embedded artificial intelligence in everyday systems. It also notes the importance of platforms, new industrial processes such as 3D printing, and decentralised peer to peer technologies that hold potential to blur traditional definitions of ownership and employment. An unresolved tension is identified between the seeming ubiquity of digital technology and a downshift in measured productivity growth. Finally, the risks of technological determinism, as well as underestimating the power of technology in the longer term, are discussed.

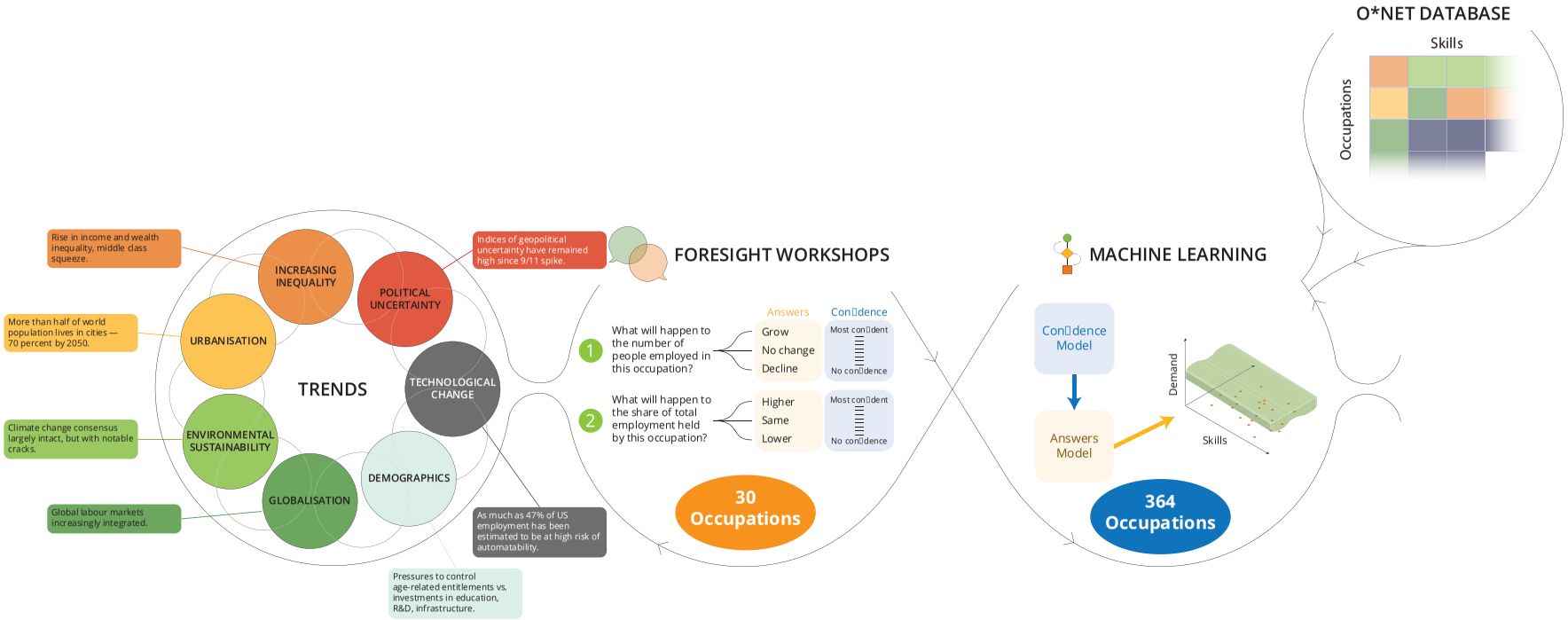


Figure 6.1: This is a holistic view of the flow of information in our research. A brief summary of the seven trends are presented which contextualise and guide discussions in the foresight workshops. Within the foresight workshops participants are presented with 30 occupations to consider. Two questions are asked for each occupation eliciting the thought leaders views on the *absolute* and *share* change in demand for that occupation in 2030. Each question also requests the uncertainty of the participant. The answers of the future demand of occupations are used as observations in our machine learning algorithm, whereby each occupation is described by 120 skills features from the O*NET database. Inferences are made on the demand of all occupations and the 120 skills contained within the O*NET database.

2. **Globalisation** This trend focuses on the effects of increasing integration of countries into the global labour market, and the emergence of new economic power-houses such as China and India. It particularly discusses the costs of globalisation – for example declining employment and wages in industries exposed to international competition – noting that the manufacturing sector has been a lightning rod for these changes (albeit in a non-uniform way). Finally it notes that both trade opportunities and challenges remain for advanced and emerging economies respectively. This is balanced with a cautionary acknowledgement of potential risks and trade barriers, culminating in a discussion over whether globalisation has structurally 'peaked'.

3. **Demographics** This trend notes that rising dependency ratios, particularly in advanced economies, increase the importance of productivity as labour inputs slow. Two impacts of an aging population are also considered: firstly, the fiscal pressure on governments to control age-related entitlements and secondly, the potential purchasing power of aging populations to influence change in sectors such as healthcare, finance, housing and entertainment. Finally, the trend identifies the potential growth in influence of Millennials as they inherit the assets of their parents. Millennials are the first generation to come of age after the arrival of digital technology and exhibit quite different consumption and work behaviour as compared to previous generations.

4. **Environmental sustainability** This trend focuses on the striking rise in consensus about the man-made nature of global warming, and associated responses. It is noted that large-scale investment in green technologies will be necessary to meet carbon reduction targets, however there is substantial structural risk of resistance to change. The wide-ranging consequences of climate change are also discussed, including less visible effects such as the potential for risk accumulation due to linkages between socio-economic and technical systems (for example climate change-induced food insecurity, linked to an increased risk of poverty and conflict). Finally, it is cautioned that

structural changes associated with developing the green economy are dependent on government policy, which currently lacks comprehensive multi-lateral coordination and can be vulnerable to political reversal.

5. **Urbanisation** This trend explores potential effects of the prediction that 70% of the world's population is expected to live in cities by 2050. Cities are magnets for high-value, knowledge-intensive industries and offer more varied consumption and employment opportunities, however, access to these opportunities is unevenly distributed across urban populations. There is also a notable disparity between the economic advantages for cities, and their surroundings, as compared to more remote areas. An increased push towards investment in 'smart cities' has been identified; leveraging information to optimise performance for example in sustainability or transport efficiency. Finally, this trend is framed by a major uncertainty in how debates about the role of fiscal policy and government activism will resolve themselves against a backdrop of high public debt ratios.

6. **Increasing inequality** This trend notes how recent decades have seen a sharp rise in income and wealth inequality, accompanied by a squeeze on the middle class and decreasing levels of economic mobility between generations. Economic distress and erosion of opportunities for people with low education have in turn created a web of social issues. These have sectoral consequences, for example rising demand for health and social care and disparities in consumption, particularly of services such as education. Factors identified as driving higher inequality include the impact of technology and globalisation, failings of the educational system, anti-competitive practices, weaknesses in corporate governance, the decline in union membership and the progressivity of the tax system. Finally it is predicted that while some of these trends may reverse in future, current levels of inequality are likely to persist in the medium-term, barring an extreme shock of some kind.

7. **Political uncertainty** This trend examines how a greater distribution of power has challenged the international system's ability to respond effectively to a host of regional and global challenges - from the spread of nuclear weapons and authoritarianism, through historical rivalries in the Middle East and Asia-Pacific, to a growing rejection of free trade and immigration. Increased uncertainty is found to have negative impacts on economic activity, which is particularly felt in sectors such as defence, finance and engineering. Finally, a discussion of policy change is raised including: elevated policy uncertainty, the weakening of institutional structures designed to promote credibility and consistency, and an increase in partisanship that can impeded effective negotiation. It is noted that even in systems designed to produce moderation, issues of marginalisation, public apathy and dissatisfaction have arisen.

6.2.2 Foresight analysis

The trends analysis was used to contextualise and guide discussions at two foresight workshops that we convened between small groups of thought leaders with domain expertise in at least one of the seven trends identified. These foresight workshops were held in Boston on 20th October 2016 with 12 participants and in London on 28th October 2016 with 13 participants.

In the second part of each of the workshops, the domain experts were presented with three sets of ten individual occupations at the 6- and 4-digit Standard Occupation Classification (SOC) level, for US and UK data respectively, and invited to debate their future prospects. They then assigned labels to the occupations (individually) according to whether they thought they would experience rising, unchanged or declining demand by 2030.[1] We define 'demand' as the share of employment in the occupation: rising demand means that the occupation will account for a larger fraction of all workers in 2030 than in 2016. The experts were also asked to record how certain/uncertain they were in making their predictions. More details on this data-gathering-process is presented in Section 6.3.3. Factsheets

[1]The occupations are those making up the US Bureau of Labor Statistics (BLS) 2010 Standard Occupation Classification (SOC).

presenting information on each of the thirty occupations (containing a list of related job titles, related industries, key skills and tasks) and their historical growth patterns were made available to the experts when making their predictions.

6.2.3 Machine learning

We used these labels allocated by the domain experts to train a machine learning classifier to generate predictions for *all* occupations, making use of a detailed data set of 120 skills, abilities and knowledge requirements against which the US Department of Labor's O*NET service scores all 4-digit occupations in the US SOC on a consistent and on-going basis, as detailed in Section 6.4.1. To maximise the performance of the algorithm, we used an active learning method, described in Section 6.4.2, whereby the second and third sets of ten occupations to be labelled were selected by the algorithm itself (intuitively, these occupations were selected to cover that part of the skills/abilities/knowledge space where the algorithm exhibited highest levels of uncertainty based on the previously labelled occupations).[2] We backed out from the model which skills, abilities and knowledge features were most associated (on their own and together) with rising or declining demand.

6.2.4 Research design and challenges

Our mixed methodology approach – making use of structured foresight and supervised machine learning techniques – was crafted to tackle the limitations in traditional qualitative and quantitative exercises. In particular, and as discussed earlier, qualitative approaches based purely on eliciting the judgments of experts are likely to be subject to human biases, while quantitative approaches based purely on trend extrapolation are likely to miss structural breaks in past trends and behaviours. By combining a machine learning algorithm with structured expert judgment we hope to have the best of both worlds.

[2]The selection of the first ten occupations presented to the experts was random: specifically, ten occupations were uniformly randomly selected, but occupations were replaced with another randomly selected occupation in the case that historical time-series were not available at least back to 1983 in the case of the US. This constraint meant that in the US we drew the occupations from 125 of the total 840 6-digit SOC codes.

Our research design is elaborate, matching the ambitious nature of our research question, but it is important to note that, as a consequence, our findings could reflect any number of assumptions. For example, the subjective judgments of one group of domain experts could be very different to another, or some parts of the O*NET data set could be more accurate characterisations of occupations than others. The provision of historical data on occupations and the main trends designed to establish a common frame of reference among experts mitigates some of these risks. However, it remains the case that predictions generated might have differed if a different group of experts had participated in the workshops or if we had used different selections of O*NET features in our model.

A separate, though related, challenge is how to evaluate our findings. As a forward-looking exercise, we might simply compare our predictions with labour market outcomes in 2030. Notwithstanding the fact that this is ten years away, a concern is that because our predictions are conditional (see above), we cannot in any straightforward way identify the source of prediction errors.

We partly tackle this by investigating the sensitivity of our findings to key features of our research design. In particular, we present predictions that used (non-parametric) trend extrapolation of employment in an occupation to label the thirty occupations in place of the experts' judgments. These data-driven labels give a baseline against which those built on our foresight exercises can be compared.

6.3 Data

6.3.1 O*NET

To derive the demand for skills, abilities and knowledge from our occupational projections, we rely on data from the O*NET, a survey produced for the US Department of Labor (Occupational Information Network (O*NET) 2017). The O*NET survey contains information on over 1000 detailed occupations, using a modified form of the Standard Occupation Classification (SOC) system. It began in 1998 and is updated on a rolling basis by surveys of each occupation's worker population as well as the assessments of job analysts.

The scope and sampling of O*NET are viewed as an improvement on its predecessor, the Dictionary of Occupational Titles (DOT) and also standard household surveys where self-reporting can result in substantial measurement errors. Reported response rates are high – at around 65% – and have been rising over time (Handel 2016). We take advantage of the 2016 O*NET to reflect most accurately the current make-up of occupations, though results from previous versions of O*NET are broadly similar.

A major strength of O*NET is that it asks many different questions about the skills, abilities, knowledge and work activities which make up occupations. Respondents/analysts are asked about the importance of a particular feature for a job (e.g. critical thinking, persuasion, manual dexterity etc.) and the level or amount of the feature required to perform it. The questions are rated on an ordinal scale which are standardised to a scale ranging from 0 to 100. We use all 120 features from the skills, abilities and knowledge categories in O*NET, designed to provide as rich a picture of occupations as possible[3]. These features are tabulated in Table 6.1.

Table 6.1: List of all O*NET features used in this study.

Type	Feature	Type	Feature
Skill	Reading Comprehension	Knowledge	Fine Arts
Skill	Active Listening	Knowledge	History and Archeology
Skill	Writing	Knowledge	Philosophy and Theology
Skill	Speaking	Knowledge	Public Safety and Security
Skill	Mathematics	Knowledge	Law and Government
Skill	Science	Knowledge	Telecommunications
Skill	Critical Thinking	Knowledge	Communications and Media
Skill	Active Learning	Knowledge	Transportation
Skill	Learning Strategies	Ability	Oral Comprehension
Skill	Monitoring	Ability	Written Comprehension
Skill	Social Perceptiveness	Ability	Oral Expression
Skill	Coordination	Ability	Written Expression
Skill	Persuasion	Ability	Fluency of Ideas
Skill	Negotiation	Ability	Originality
Skill	Instructing	Ability	Problem Sensitivity
Skill	Service Orientation	Ability	Deductive Reasoning
Skill	Complex Problem Solving	Ability	Inductive Reasoning
Skill	Operations Analysis	Ability	Information Ordering

[3]According to O*NET, skills represent developed capacities which facilitate learning or the more rapid acquisition of knowledge; abilities are enduring attributes of the individual which influence performance, and knowledge refers to organised sets of principles and facts applying in general domains.

Table 6.1: List of all O*NET features used in this study.

Type	Feature	Type	Feature
Skill	Technology Design	Ability	Category Flexibility
Skill	Equipment Selection	Ability	Mathematical Reasoning
Skill	Installation	Ability	Number Facility
Skill	Programming	Ability	Memorization
Skill	Operation Monitoring	Ability	Speed of Closure
Skill	Operation and Control	Ability	Flexibility of Closure
Skill	Equipment Maintenance	Ability	Perceptual Speed
Skill	Troubleshooting	Ability	Spatial Orientation
Skill	Repairing	Ability	Visualization
Skill	Quality Control Analysis	Ability	Selective Attention
Skill	Judgment and Decision Making	Ability	Time Sharing
Skill	Systems Analysis	Ability	Arm-Hand Steadiness
Skill	Systems Evaluation	Ability	Manual Dexterity
Skill	Time Management	Ability	Finger Dexterity
Skill	Management of Financial Resources	Ability	Control Precision
Skill	Management of Material Resources	Ability	Multilimb Coordination
Skill	Management of Personnel Resources	Ability	Response Orientation
Knowledge	Administration and Management	Ability	Rate Control
Knowledge	Clerical	Ability	Reaction Time
Knowledge	Economics and Accounting	Ability	Wrist-Finger Speed
Knowledge	Sales and Marketing	Ability	Speed of Limb Movement
Knowledge	Customer and Personal Service	Ability	Static Strength
Knowledge	Personnel and Human Resources	Ability	Explosive Strength
Knowledge	Production and Processing	Ability	Dynamic Strength
Knowledge	Food Production	Ability	Trunk Strength
Knowledge	Computers and Electronics	Ability	Stamina
Knowledge	Engineering and Technology	Ability	Extent Flexibility
Knowledge	Design	Ability	Dynamic Flexibility
Knowledge	Building and Construction	Ability	Gross Body Coordination
Knowledge	Mechanical	Ability	Gross Body Equilibrium
Knowledge	Mathematics	Ability	Near Vision
Knowledge	Physics	Ability	Far Vision
Knowledge	Chemistry	Ability	Visual Color Discrimination
Knowledge	Biology	Ability	Night Vision
Knowledge	Psychology	Ability	Peripheral Vision
Knowledge	Sociology and Anthropology	Ability	Depth Perception
Knowledge	Geography	Ability	Glare Sensitivity
Knowledge	Medicine and Dentistry	Ability	Hearing Sensitivity
Knowledge	Therapy and Counseling	Ability	Auditory Attention
Knowledge	Education and Training	Ability	Sound Localization
Knowledge	English Language	Ability	Speech Recognition
Knowledge	Foreign Language	Ability	Speech Clarity

Our implementation strategy departs from Frey and M. A. Osborne's (2017) study of automation in that it relies on O*NET's 'importance' rating. Analyses of O*NET data suggest that there is substantial overlap between the importance and level ratings, so this modelling choice does not lead to vastly different predictions

in practice. Critically, the importance rating is available for all combinations of features and occupations. This is in marked contrast to the level rating for which O*NET recommends suppressing a large number of estimates on account of their low precision. This problem is most serious for knowledge features; to implement O*NET's recommendations in full would entail removing occupations equivalent to 89% of total US employment. Similarly, the scales and anchor points used to construct the level ratings have been criticised for their complexity which may affect the reliability of some ratings (Handel 2016).

In the remainder of this report, we will use x to represent the vector of length 120 containing these variables. We also include in our workshop factsheets the occupation description and five common job title examples for the occupation, also taken from O*NET.

6.3.2 Employment Microdata

To form yearly estimates of employment by occupation and industry for our workshop factsheets, we used 1983 – 2015 data from the Current Population Survey's (CPS) Annual Social and Economic Supplement (ASEC) from the Integrated Public Use Microdata Series (IPUMS) provided by the Minnesota Population Center (M. King et al. 2010). We used an Integrated Public Use Microdata Series (IPUMS)-provided best-guess harmonization of occupation codes over time to 1990 Census occupation code. We then crosswalked these codes to 6-digit US SOC 2010 codes.

For industry, we presented Current Population Survey (CPS)-derived estimates of occupation employment by industry in 2015, harmonized in the same way. We used the most granular common level of North American Industry Classification System (NAICS) 2012 code available for that occupation (either the four-, three-, or two-digit level).

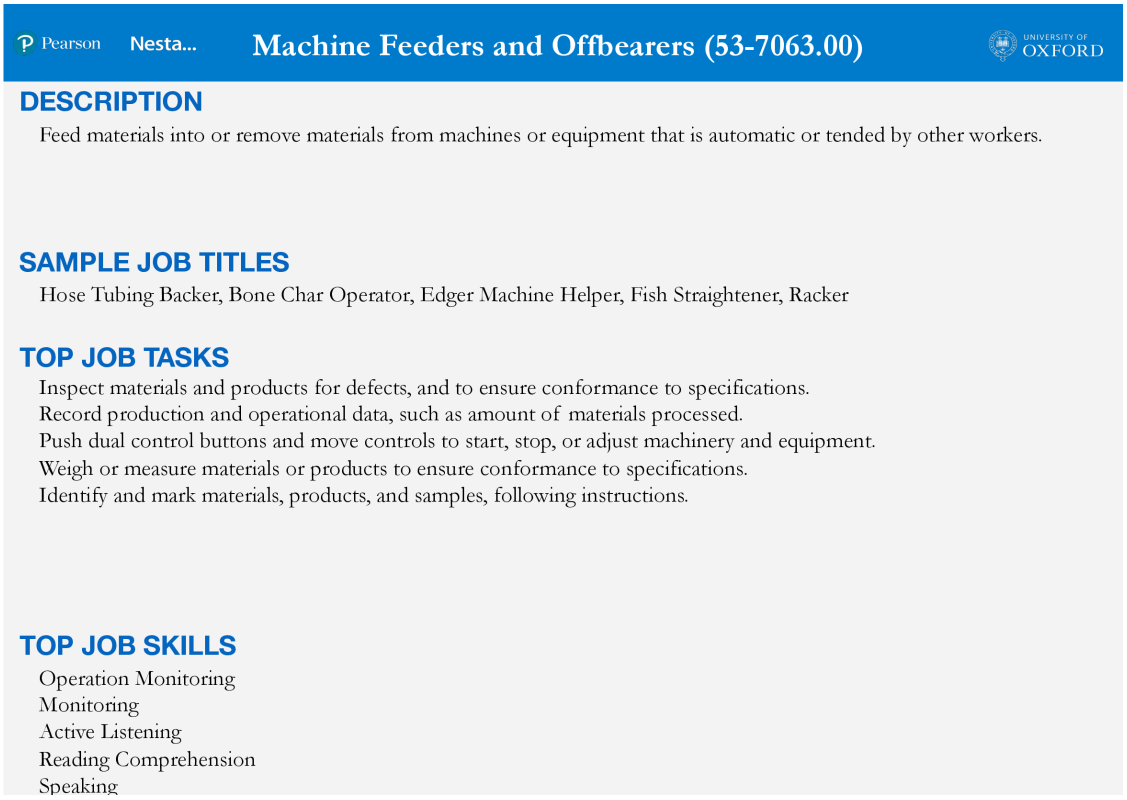
The US employment results from our machine learning classifier are weighted using the May 2015 Occupational Employment Statistics from the Bureau of Labor Statistics (U.S. Bureau of Labor Statistics 2015).

6.3.3 Workshop-Generated Data

To collect labels to train our machine learning classifier of future demand for occupations, we held two expert foresight workshops in Boston and London in October 2016. Each workshop brought together a diverse group of 12-13 experts from industry, government, academia, and the social sector. Our experts were instructed to consider the net impact on the workforce occupation composition of all the trends discussed above, guided by presentation of our trends analysis.

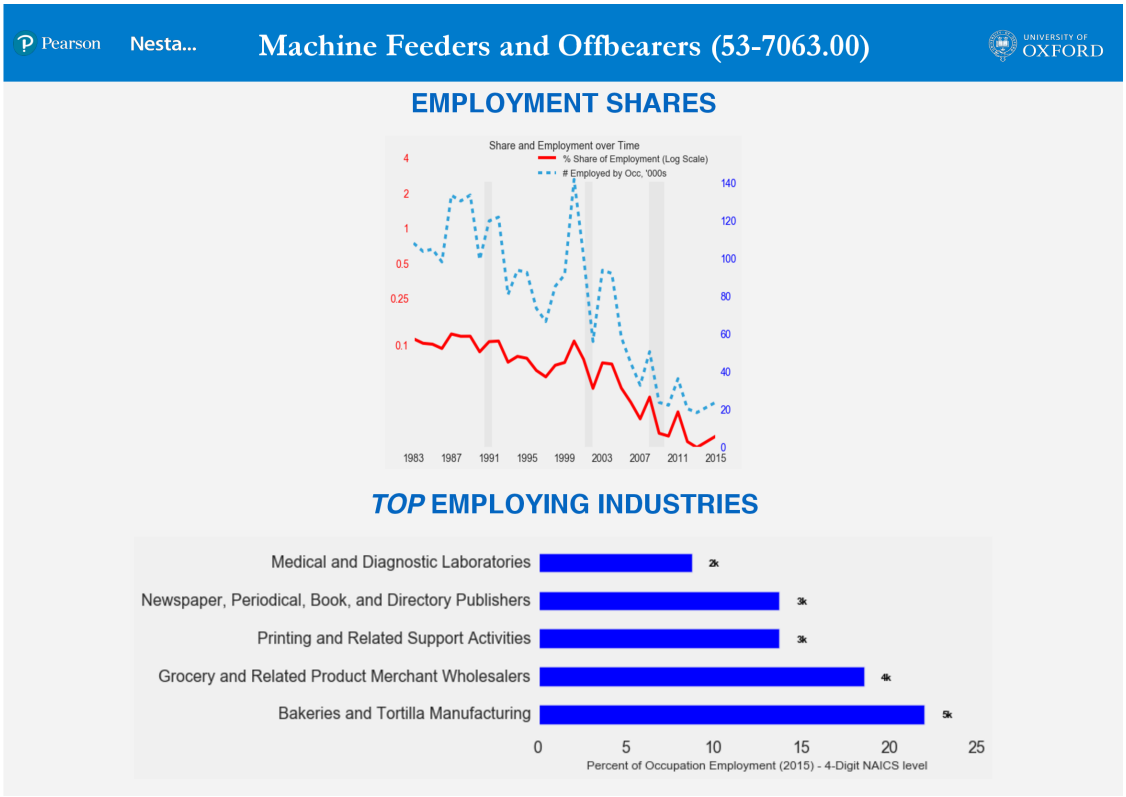
Over the course of the workshop, the group participated in three prediction sessions. In each session, the participants viewed the information described above for ten occupations, displayed on two slides ("factsheets") per occupation, as shown in Figures 6.2 and 6.3.

Figure 6.2: Factsheet for US occupation *Machine Feeders and Offbearers*



Note that the factsheets presented time-series plots of the occupation to participants, such that they could form their predictions with proper historical context.

Figure 6.3: *Statistics* Factsheet for US occupation *Farm workers*



After viewing the occupation descriptions the group was directed to an online form to answer two three-choice questions:

1. *What will happen to the share of total employment held by this occupation?* {Higher share, Same share, Lower share}

2. *What will happen to the number of people employed in this occupation?* {Grow, No change, Decline}

With only a three-point scale, it was important to consider both employment share and absolute employment levels so as to allow a fuller expression of judgments of future demand. For example, only knowing that an occupation will grow slower than the workforce as a whole says nothing about whether it will add or shed jobs. [4] In this event, the US workshop was of the view – albeit with significant

[4]Consider 'engineers' and 'metal workers and plastic workers' – two occupation groups which the US Bureau of Labor Statistics (BLS) projects will decline in share between 2014 and 2024. Over this period, 'metal workers and plastic workers' are projected to lose 99,000 jobs whereas 'engineers' are projected to add 65,000 jobs.

differences in opinion across individuals – that total employment would grow over the prediction horizon, consistent with the historical pattern.

Given the inherent difficulties in making long-term predictions, our workshop participants were also asked to provide a 0-9 ranking of how *certain* they were in their answer, with *0* representing *not certain at all*, and *9* representing *completely certain.* They were also given a space to provide freeform thoughts they felt were necessary to qualify their answers.

After the group submitted their answers using the online form, an experienced foresight workshop facilitator reviewed the responses with the group. After the group debated their perspectives during a half-hour session, the group was then allowed to change their answers, after which the workshop moved onto the next set of ten occupations.

The first ten occupations presented to participants were selected randomly. For the second and third rounds, the respective batches of ten occupations were selected so as to be maximally informative for the machine learning model in light of the answers previously gathered from participants. In a way, the participants could be seen as teaching the machine learning algorithm throughout the course of the day, with the algorithm able to respond to the information from participants by proposing a prioritised list of further questions. This process will be described formally in Section 6.4.2.

6.4 Methodology

Our methodology uses the foresight exercises described in Section 6.3.3 as training data for a machine learning model. The primary goal of the model is to learn a function $f(x)$ that maps from 120 O*NET variables x (capturing skills, knowledge and abilities) to future occupational demand. In this framework, the ith occupation is considered a point, $x^{(i)}$, in 120-dimensional skills-/knowledge-/ability-space, whose associated demand is $f(x^{(i)})$. Our approach is built on an expectation that demand should vary smoothly as a function of skills, knowledge and abilities: that is, if

two occupations i and j have similar O*NET variables, $x^{(i)} \simeq x^{(j)}$, we expect their associated demands to be similar, $f(x^{(i)}) \simeq f(x^{(j)})$.

We choose to model f with a Gaussian Process (GP). The GP gives a flexible, non-linear, function class suitable for the complex interactions (for instance, complementarities) that we expect between variables and demand.

This model is trained on the dataset created from the workshop, qualitatively described in Section 6.3.3. The dataset contains ternary labels for each absolute and share demand question answered for each occupation surveyed by each individual expert. This approach permits the diversity of views within the group to be captured within the model. The participant labels are modelled as conditionally independent given the latent function $f(x)$. The dependence amongst the group's labels induced by discussion is modelled through this shared latent function.

Resilience to uncertainty is crucial to our exercise. Not only did our workshops gather observations from individual participants with explicit representations of their uncertainty, the model must also try to fuse observations from the diverse range of opinions produced by our domain experts. Our model choice is informed by the probabilistic foundations of the GP, which give it a coherent way to reason about uncertainty. As such, we expect our model to give an honest representation of the trends that can be inferred from noisy participant labels.

Our model is put to work on a variety of tasks. In particular, it is the basis of our means of selecting occupations to be labeled through active learning (Section 6.4.2), and interpreting the patterns discovered by the model is the basis of our assessment of the importance of O*NET variables to future demand (Section 6.4.3). Finally, we use the model to predict future occupations, defined as hotspots of high demand that are not associated with an existing occupation (Section 6.4.4).

To benchmark the efficacy of our foresight exercise, we also use GPs to perform extrapolation out to the year 2030 of past employment trends (Section 6.4.5). These extrapolations provide alternatives to the labels produced in the workshops, and provide results that do not rely on the subjective judgments of the experts.

6.4.1 Heteroscedastic ordinal regression

We use the GP-HOR model from Chapter 3 as a core component to build our future of skills model. As there are two questions being asked, one relating to share change, and the other to absolute change, two different types of observations are made. Subscripts s and a will represent share and absolute change, respectively. The latent function $f(\boldsymbol{x})$ of our GP-HOR model represents the change in demand in absolute employment. The relative change in share can be represented by dividing the absolute change in demand by the total size T of the workforce in 2030. That is, the second of the two workshop questions provides an observation f/T_m, where T_m is an unknown positive value to be inferred for participant m and $T_m \in \mathbb{R}^+$.

For each question, the model ingests ternary-valued labels from one of the sets {Lower share, Same share, Higher share} and {Decline, No change, Grow}, for relative share of employment and absolute change in employment, respectively. Let $y_s \in \{0, 1, 2\}$ and $y_a \in \{0, 1, 2\}$ represent share change and absolute change, respectively; with the following mappings to the above meanings:

$$y_s : \{0, 1, 2\} \to \{\text{Lower share, Same share, Higher share}\} \tag{6.1}$$

$$y_a : \{0, 1, 2\} \to \{\text{Decline, No change, Grow}\} \tag{6.2}$$

For clarity and interpretability of the model, the decision was made to report results for binary-valued labels $y_{out:s} \in \{0, 1\}$, namely, {Decreasing demand, Increasing demand}, which are used as the foundation of the analysis.

Recall that participants describe how confident they are using a choice from an ordinal list $\{0, \ldots, 9\}$, we represent this with the variable y_c for both absolute and share change observations:

$$y_c : \{0, ..., 9\} \to \{\text{No confidence, ..., Most confident}\} \tag{6.3}$$

Since the value of the noise is hence an ordinal variable as well, we build a secondary ordinal regression model to predict the ordinal value of noise at different points in feature space. Just as in the GP-HOR model developed Chapter 3 the secondary latent function $g(\boldsymbol{x})$ represents the level of certainty of the participant.

The relationship between the share and absolute change ternary-valued labels $(y_s(\boldsymbol{x}), y_a(\boldsymbol{x}))$ and $f(\boldsymbol{x})$ and $y_c(\boldsymbol{x})$ is shown below:

$$y_s(\boldsymbol{x}) = \begin{cases} 0, & -d_0 < \frac{f(\boldsymbol{x})}{T_m} + \epsilon_q(y_c(\boldsymbol{x})) \leq -d_1; \\ 1, & -d_1 < \frac{f(\boldsymbol{x})}{T_m} + \epsilon_q(y_c(\boldsymbol{x})) \leq d_1; \\ 2, & d_1 < \frac{f(\boldsymbol{x})}{T_m} + \epsilon_q(y_c(\boldsymbol{x})) \leq d_2; \end{cases} \tag{6.4}$$

$$y_a(\boldsymbol{x}) = \begin{cases} 0, & -d_0 < f(\boldsymbol{x}) + \epsilon_q(y_c(\boldsymbol{x})) \leq -d_1; \\ 1, & -d_1 < f(\boldsymbol{x}) + \epsilon_q(y_c(\boldsymbol{x})) \leq d_2; \\ 2, & d_2 < f(\boldsymbol{x}) + \epsilon_q(y_c(\boldsymbol{x})) \leq d_3; \end{cases}, \tag{6.5}$$

where $[d_0, ..., d_2] \in \mathbb{R}^{+4}$ provide symmetric latent value thresholds for different ordinal values of $y_s(\boldsymbol{x}), y_a(\boldsymbol{x})$ and the boundary value of the threshold is $d_0 = \infty$. $\epsilon_q(y_c(\boldsymbol{x}))$ is the observational noise, note the dependence on confidence label $y_c(\boldsymbol{x})$.

Similarly the relationship between the confidence many-valued labels $(y_c(\boldsymbol{x}))$ and $g(\boldsymbol{x})$ is shown below:

$$y_c(\boldsymbol{x}) = \begin{cases} 0, & e_0 < g(\boldsymbol{x}) + \epsilon_c \leq e_1; \\ \vdots & \vdots \\ 9, & e_9 < g(\boldsymbol{x}) + \epsilon_c \leq e_{10}. \end{cases}, \tag{6.6}$$

where $[e_0, ..., e_{10}] \in \mathbb{R}^{11}$ are latent value thresholds for different ordinal values of y_c and the boundaries of the threshold are $e_0 = -\infty$ and $e_{10} = \infty$. ϵ_c is the observational noise for the confidence.

We assume the observational noises are drawn from a Gaussian distributions as follows:

$$\epsilon_q(y_c(\boldsymbol{x})) \sim \mathcal{N}\left(0, \sigma_{\text{q-noise}}(y_c(\boldsymbol{x}))\right) \tag{6.7}$$

$$\epsilon_c \sim \mathcal{N}\left(0, \sigma_{\text{c-noise}}\right), \tag{6.8}$$

where $\sigma_{\text{q-noise}}(y_c(\boldsymbol{x}))$ and $\sigma_{\text{c-noise}}$ are standard deviation of the observational noise of the question and confidence, respectively. The standard deviation of the question

observational noise is defined as:

$$\sigma_{\text{q-noise}}(y_c) = \begin{cases} \sigma_{q:0} & y_c = 0 \\ \vdots & \vdots \\ \sigma_{q:9} & y_c = 9 \end{cases}, \tag{6.9}$$

where we assume that the noise standard deviation associated with each question observation is an affine transformation of the chosen value: $[\sigma_{q:0}, ..., \sigma_{q:9}] = [\alpha + (i_{max} - i)\beta | i = 0, ..., 9]$ for $i_{max} = 9$ with hyperparameters $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$. Hyperparameters that are strictly positive have positivity enforced by passing that variable through a log transform.

Likelihood Marginalising out the observational noise terms ϵ_q and ϵ_c our likelihood for the absolute change observation can be taken directly from our GP-HOR model of Chapter 3 which is:

$$p(y_{\text{a}} \mid f(\boldsymbol{x}_i), y_c) = \Phi(u^i_{a:1}) - \Phi(u^i_{a:2}), \tag{6.10}$$

where i is the ith absolute change observation, $u^i_{a:1} = \frac{d_{y^i_a} - f(\boldsymbol{x}_i)}{\sigma_{\text{q-noise}}(y_c)}$, $u^i_{a:2} = \frac{d_{y^i_a - 1} - f(\boldsymbol{x}_i)}{\sigma_{\text{q-noise}}(y_c)}$ and $\Phi(u) = \int_{-\infty}^{u} \mathcal{N}(\xi, 0; 1) d\xi$

The likelihood for share change is similar to $p(y_{\text{a}} \mid f(\boldsymbol{x}_i), y_c)$ but has the latent function $f(\boldsymbol{x})$ divided by hyperparameter T_m:

$$p(y_{\text{s}} \mid f(\boldsymbol{x}_j), y_c) = \Phi(u^j_{s:1}) - \Phi(u^j_{s:2}) \tag{6.11}$$

where j is the jth share change observation, $u^j_{s:1} = \frac{d_{y^j_s} - f(\boldsymbol{x}_j)/T_m}{\sigma_{\text{q-noise}}(y_c)}$, $u^j_{s:2} = \frac{d_{y^j_s - 1} - f(\boldsymbol{x}_j)/T_m}{\sigma_{\text{q-noise}}(y_c)}$

The likelihood for confidence is taken from Section 3.3.3:

$$p(y_{\text{c}} \mid g(\boldsymbol{x}_k)) = \Phi(u^k_{c:1}) - \Phi(u^k_{c:2}) \tag{6.12}$$

where k is the kth confidence observation, $u^k_{c:1} = \frac{e_{y^k_c} - g(\boldsymbol{x}_k)}{\sigma_{\text{c-noise}}}$, $u^k_{c:2} = \frac{e_{y^k_c - 1} - g(\boldsymbol{x}_k)}{\sigma_{\text{c-noise}}}$ The total likelihood of all observations from both questions

$$p(\mathcal{D} \mid f(\boldsymbol{x}), g(\boldsymbol{x})) = \prod_{i=1}^{n} \left[p(y_a \mid f(\boldsymbol{x}_i), y_c) p(y_c \mid g(\boldsymbol{x}_i)) \right] \left[p(y_s \mid f(\boldsymbol{x}_i), y_c) p(y_c \mid g(\boldsymbol{x}_i)) \right], \tag{6.13}$$

where $\mathcal{D}$ is the set of all observations, n are the number of occupations for which observations were elicited.

Prior The prior for both latent functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are GPs with zero-mean function and a Squared Exponential (SE) Automatic Relevance Determination (ARD) kernel function.

Inference We employ the variational Gaussian approximation for inference, as described in Section 3.3.3. This assumes multivariate Gaussian distributions $q(\boldsymbol{f})$ and $q(\boldsymbol{g})$ can be used to approximate the posterior distributions $p(\boldsymbol{f}|\mathcal{D}_s)$ and $p(\boldsymbol{g}|\mathcal{D}_c)$, respectively. Implementation is carried out using the gpflow python package.

The main alteration of the inference method detailed in Section 3.3.3 is that the score likelihood term has two components, namely, absolute (Eq. 6.10) and share (Eq. 6.11). The share likelihood component contains a variable T_m which the absolute component does not. In the inference T_m is treated as hyperparameter to be learnt. Within gpflow the absolute and share demand likelihood components are called as separate observations with a binary variable toggling between the two likelihood functions.

During the inference phase the demand and confidence posterior were trained independently due to $p(y_c^i)$ being assumed to be a delta function around the observed confidence label.

Prediction As stated before for clarity and interpretability presentation of results were chosen to be through a binary-valued output label, which uses the same likelihood from equation 6.11 but setting $d_1 = 0$. This in effect removes the *same share* label. The posterior of the binary-valued share change ($y_{out:s}$) conditional on a new point $\boldsymbol{x}_*$ and new confidence noise label y_c^*:

$$q(y_{out:s}^* \mid \boldsymbol{x}_*, y_c^*, \mathcal{D}) = \int df_* \; p(y_{out:s}^* \mid f_*, y_c^*) q(f_* \mid \boldsymbol{x}_*, \mathcal{D}) \tag{6.14}$$

The posterior of the new confidence noise label y_c^* is:

$$q(y_c^* \mid \boldsymbol{x}_*, \mathcal{D}) = \int dg_* \; p(y_c^* \mid g_*, y_c^*) q(g_* \mid \boldsymbol{x}_*, \mathcal{D}) \tag{6.15}$$

This leads to the posterior of the share change conditional only on a new point $\boldsymbol{x}_*$, whereby the confidence noise label is marginalised out:

$$q(y^*_{out:s} \mid \boldsymbol{x}_*, \mathcal{D}) = \sum_{y_c} [q(y^*_{out:s} \mid \boldsymbol{x}_*, y^*_c, \mathcal{D}) q(y^*_c \mid \boldsymbol{x}_*, \mathcal{D})] \tag{6.16}$$

The derivative of 'Increasing demand of an occupation' with respect to the occupation features is made possible by the use of automatic differentiation, a feature of Google's Tensorflow Python package (Martin Abadi et al. 2015), which provides the framework to GPflow. Mathematically the derivative of 'Increasing demand of an occupation' with respect to the occupation features is:

$$\left. \frac{\partial q(y^*_{out:s} \mid \boldsymbol{x}_*, \mathcal{D})}{\partial \boldsymbol{x}_*} \right|_{y^*_{out:s}=1} \tag{6.17}$$

6.4.2 Active learning

As described in Section 6.3.3, our foresight workshops required choosing which occupations were to be presented to participants. We introduced the use of a machine learning model to automate this choice. The machine learning model used was a reduced form of the model described in Section 6.4.1 to ensure 'real-time' performance.[5] In particular, given that our goal is to predict demand, and that our model is able to provide estimates of the uncertainty of demand in all occupations, we took the natural option of *uncertainty sampling.* Uncertainty sampling ranks the set of all occupations from high- to low- uncertainty, and chooses the highest-ranked for labelling by participants. The motivation for the approach is the expectation these labels should be most informative about demand overall: their observation should lead to a large reduction in total uncertainty. The active learning approach is one that aims to interactively acquire data so as to provide the greatest confidence in resulting predictions.

[5]Observations consisted only of the ternary-valued labels and participant uncertainty was not incorporated. Reflecting the dataset used in (Frey and M. A. Osborne 2017) only nine features were used, namely: 'Originality', 'Systems Evaluation', 'Hearing Sensitivity', 'Arm-Hand Steadiness', 'Learning Strategies', 'Oral Comprehension', 'Social Perceptiveness', 'Manual Dexterity', 'Problem Sensitivity'.

6.4.3 Assessing feature importance

One of the core goals of this work is to assess the significance of the 120 O*NET features to future demand, and thereby inform skills policy decisions.

First, however, our research question needs further clarification: what exactly does it mean for a feature to be *important* to demand? We propose two primary desiderata for a scheme to measure importance:

1. An important feature must be clearly predictive of demand; and
2. An increase in an important feature must lead to a strong increase in demand.

We also propose two secondary desiderata for a scheme to measure importance:

1. it must be able to uncover non-linear interactions between features; and
2. it must be able to capture complementarities between features; we wish to discover features whose importance is contingent on the values of other features.

One approach to assessing feature importance is *feature selection* (Guyon and Elisseeff 2003). Feature selection is a broad and well-studied topic, and aims to choose those features that are most informative of the function. In our context, it might be thought that the ranking of O*NET features selected through such a scheme is a means of ranking their importance to demand. We suggest that the bulk of methods of feature selection give, at best, an insufficient guide to importance, and, at worst, actively misleading: feature selection does not address the second of our primary desiderata. That is, determining that a feature is highly informative gives no sense of the *sign* of the relationship between feature and demand. Most feature selection adopts an information-theoretic approach that would not distinguish between a feature x_1, for which $f(x) \simeq \alpha - x_1$, and a feature x_2, for which $f(x) \simeq \beta + x_2$ (α and β being some parameters). Both x_1 and x_2 are highly informative of demand. However, a skills policy that result in broad increases in x_1 would lead to harmful outcomes for occupations; x_2, the converse. Note also

that, for our complex, non-linear, function f, relationships with x_i are unlikely to be as simple as that described above for x_1 and x_2. Another feature, x_3, may give $f(x) \simeq \cos(x_3)$: while, again, x_3 is highly informative of demand, it is unclear whether it is *important*: for some occupations (values of x_3), x_3 will have very different significance from that for other occupations. To be explicit, the *Automatic Relevance Determination* approach (CE Rasmussen and C. Williams 2006) often used for embedded feature selection for GPs is inappropriate for our ends. It provides only a description of the informativeness of features, rather than their importance.

A second approach to managing features is *dimensionality reduction*, which would involve projecting the data into a lower-dimensional space. To give an example, dimensionality reduction can be achieved by the ubiquitous technique of Principal Component Analysis (Pearson 1901). Dimensionality reduction on x can certainly be used to discover that certain features co-vary. This, of course, is not the same as discovering that the two are similarly important to demand. More sophisticated uses of dimensionality reduction, that include the values of $f(x)$ itself, can be used to discover relationships between O*NET features and demand. However, dimensionality reduction, in considering combinations of features, will fail to satisfy our first desideratum. That is, in mixing features together, dimensionality reduction will fail to uncover clear and interpretable relationships to increasing demand.

We make two complementary proposals for assessing feature importance, and ultimately present results from each.

Pearson correlation

We first consider a direct means of achieving our primary two desiderata, while ignoring the secondary desiderata. This metric of the importance of a feature, which we abbreviate as *Pearson correlation*, is the employment-weighted Pearson correlation coefficient[6] between our model's predictive mean for demand and the feature. More precisely, let the posterior mean for the latent demand feature

[6]Note that the Pearson correlation coefficient is often used for feature selection (as a filter), an exception to the inappropriate information-theoretic methods of feature selection we generically describe above.

for the ith occupation be $m(x^{(i)})$. We can then define the Pearson correlation value for the nth O*NET feature as

$$\mathrm{PC}(n) := \frac{\sum_{i=1}^{I} w(i)\Big(m(x^{(i)}) - \mathbb{E}\Big(m(x)\Big)\Big)\Big(x_n^{(i)} - \mathbb{E}\Big(x_n\Big)\Big)}{\sigma\Big(m(x)\Big)\sigma\Big(x_n\Big)}, \tag{6.18}$$

where (for any function $g(x)$) we define the employment-weighted expectation and variance

$$\mathbb{E}\Big(g(x)\Big) := \sum_{i=1}^{I} w(i) g(x^{(i)}) \quad \text{and} \tag{6.19}$$

$$\sigma\Big(g(x)\Big)^2 := \mathbb{E}\Big(g(x)^2\Big) - \mathbb{E}\Big(g(x)\Big)^2, \tag{6.20}$$

I is the total number of occupations, and $w(i)$ is the fraction of total employment within the ith occupation.

Pearson correlation measures the *linear* relationship between mean value of demand and a feature. As such, it gives the sign of clear relationships, but satisfies neither of our secondary desiderata. One consequence of linearity is that the Pearson correlation may place low weight on features that are linked to high demand only for a small number of occupations. Another consideration is that it does not consider the covariance of latent demand, only the mean. This has the potential to place importance on occupations which have high uncertainty from the model. Nonetheless, the features that it does highlight will unquestionably be important: if a strong positive linear interaction exists, it should certainly influence our resulting skills policy. As such, we would consider Pearson correlation to give a sufficient but not necessary condition for importance.

Average derivative and feature complementarity

Our second proposal gives a means of satisfying our secondary desiderata, while perhaps weakening the case for the first of the primary desiderata. The *average derivative*, as described in (Baehrens et al. 2010), is for the nth feature simply

$$\mathrm{AG}(n) := \mathbb{E}\left(\frac{\partial m(x)}{\partial x_n}\right), \tag{6.21}$$

using the employment-weighted expectation defined in (6.19).

By way of interpretation, the derivative measures the expected increase in demand for a unit increase in a particular feature (for instance, as a result of a policy intervention). By averaging over all occupations, we get a sense of the aggregate increase in demand as a result of this increase in a feature. The average derivative gives an interpretable notion of sign: it can clearly distinguish positive from negative relationships with demand.

The first advantage of this metric relative to the marginal correlation is that it is sensitive to non-linearities in the data, addressing the first of our secondary desiderata. While the derivative gives a linear approximation to demand, it is only a *locally* linear approximation. By considering the approximation at all points (occupations) in skills-/knowledge-/abilities-space, we are able to better measure relationships that have different slopes at different regions of the space.

This ability to manage non-linearity also enables the average derivative to capture the importance of features whose significance is conditional on the values of other features. For instance, Fine Arts is very important to Artists, but less important for occupations with differing skills profiles. This is achieved through reporting the derivative averaged over subsets of occupations: for instance, those that fall within a major occupational grouping[7]. For a non-linear function, the average derivative for a feature may be substantively different over one region of space (occupational grouping) than for another.

For each subset of occupations, we will highlight those features with both large positive and large negative average derivatives. Those with large positive derivatives we say are *complementary* to the occupational group (increasing such a feature increases demand), whereas those with large negative derivatives we say are *anti-complementary* (increasing such a feature decreases demand). It is also of interest to speak of complementarities between features, rather than simply of the complementarity of a feature to an occupational group. To do so, we must find some way of singling out the features associated with the occupational group. Those features that are large on average for an occupational group are, speaking roughly,

[7]To be precise, this entails only redefining the expectations above as employment-weighted sums over that subset of occupations, rather than over all occupations.

those that are most exceptionally significant, and hence most characteristic, of the occupational grouping. As such, we will define features with large positive average derivatives to be *complementary* to those characteristic (large-valued) features[8].

The key drawback of the average derivative approach is that the averaging itself may occlude signal. As an example, if $f(x) \simeq 10^{10}\cos(x_n)$, there are many points for which the average derivative would be very large: for instance, all those points immediately to the left of a peak, $\{x_n = -\epsilon + 2\pi n; \forall \text{ small positive } \epsilon \text{ and integer } n\}$. This x_n is a feature that is unlikely to be useful in policy: it has an equal number of points (occupations) with very large negative derivative. Increasing x_n would be very harmful to all such occupations. Nonetheless, if the chosen samples (occupations) contain even one more point of large positive derivative than large negative derivative, x_n may have high average derivative. This drawback leads to us regarding the average derivative as a necessary but not sufficient condition for importance; rendering it complementary to the marginal correlation.

To slightly ameliorate the problems of averaging, we calculate the empirical distribution of the derivative for each feature – particularly, its mean and standard deviation. We then exclude features whose derivative has a standard deviation that is in the top 97th percentile of the distribution. The rationale is that the influence of these features is very 'noisy' (e.g. within a particular occupational grouping, demand might strongly increase for some occupations and strongly decrease for other occupations) and hence unlikely to be a reliable basis upon which to design policy. Here we have implicitly taken a conservative view of the potential of skills

[8]Our definition of complementarity is loosely related to that used by economists: two features are complementary if the marginal value product of one is increasing in the level of the second. This definition fails to meet our needs. The first problem is that the definition makes no accommodation for location in feature space. The economics definition is a statement about the second-order derivative of the function with respect to the two features being positive; for an arbitrary function, as may be learned by our flexible non-parametric model, the second-order derivatives may be in very different regions of space. The second, related, problem with the definition is that it may lead to highlighting feature combinations which, even if the second-order derivatives are positive and constant across space, are actively harmful. As a simple example, for a bivariate quadratic function with positive-definite Hessian (a convex bowl), an occupation on the wrong side of the critical point (the minimiser, or the location of the bottom of the bowl) would see its demand decreased with increases in either or both of the features. Our means of assessing complementarity, however, will more correctly identify the differing importances of feature combinations at any point in feature space.

policy to precisely affect targeted occupational groups. However, there is a trade-off in the selection of the threshold of exclusion. Any non-linearity in demand will result in the derivative varying over x, thereby increasing the standard deviation. Note that features excluded under this scheme are excluded only for consideration by the average derivative: the recommendations of this metric cannot be trusted for these features. The excluded features are used as normal in every other facet of our modelling, as in producing occupation-level and aggregate predictions of demand.

6.4.4 New occupations

We define a potential new occupation as a combination of skills, knowledge and abilities that is likely to see high future demand, but is not associated with an existing occupation. To forecast where new occupations might emerge, we simply optimise the posterior mean for the latent demand variable, $m(x)$ as a function of x. More precisely, we start by randomly selecting 50 current occupations as our starting points of high demand occupations. We run local optimisers (Limited-memory Broyden–Fletcher–Goldfarb–Shannon (L-BFGS), observing box-constraints, as per (Byrd et al. 1995)) initialised at each of the 50 occupations. This will return 50 local optimisers of $m(x)$: points $x^{\dagger}$ in skills, knowledge and abilities space that are associated with high demand. Some of these optimisers will converge to (near-)identical existing occupations whereas many will converge to (near-)identical sets of novel features. Beginning with a single such optimiser, $x^{\dagger,0}$, we add each successive optimiser unless it is closer, in the 2-norm sense, than a preset threshold ($\epsilon = 0.1$) to either: an existing occupation, or; a previously included optimiser. This procedure will return a list of hotspots of demand (local optimisers) $\{x^{\dagger,i}; \forall i\}$, of length that may be different for different mean functions. Each hotspot can be interpreted through returning its vector of skills, knowledge and abilities values $x^{\dagger,i}$, as well as the list of occupations which are closest to it.

6.4.5 Trend extrapolation

The trend extrapolation work explained below was primarily carried out Justin Bewshew and it is included for completeness.

As an alternative to the labels from our foresight workshop participants, we use extrapolation of historical employment trends to give a more data-driven alternative to predicting demand in the year 2030. Specifically, we use a GP to regress UK and US (for the years we have) employment as a function of years, for both absolute employment values and share employment values, projecting forward to 2030.

Each occupation is modelled separately using a GP. A Matérn Covariance with parameter $\nu = 3/2$ is used for each GP; this ensures sufficiently smooth extrapolation without losing the structure in the trends. To control the characteristic scale on which the GP varies (a *length scale*), we assign a prior that centres the value of the length scale such that data points from the last year in our employment series will influence employment numbers in 2030. The GP is modelled over the log of the employment numbers.

Absolute employment numbers are modelled directly from historical data using a GP. The share values are slightly more involved. First, we model the total workforce, T, as a function of time and use our model to extrapolate that forward to 2030. We then divide the extrapolated total workforce values by the absolute values to give the share value predictions.

We use these trend extrapolations to compute the probability of the trend being higher demand, the same demand or lower demand between 2015 and 2030. The probability of being high is taken as the total positive difference above 2 standard deviations, low is taken as the total negative difference below 2 standard deviations and the same otherwise.

To generate a data set equivalent to our workshop response we sample from these extrapolated probabilities. A three-parameter multinomial distribution is sampled using the probabilities of higher, same and lower. For each occupation we sample 12 values and use these as the participant responses. A noise value of 0 is assigned for every data point.

These values are then fed into the GP-HOR model and the corresponding probabilities are computed as described in Section 6.4.1.

6.5 Results

We present below our main results, and their interpretation, for the US economy. The first analysis we present relates to the share of employment in 2030 by occupation, and our model's outputs are the probabilities of that share being greater than at present. Below, we will informally use 'increased demand', 'future demand' (and, at times, simply 'demand') as a shorthand for 'increased share of employment in 2030'.

6.5.1 Occupations

The primary outputs of our model are the probabilities of each occupation experiencing a rise in workforce share (that is, increased demand). These occupation-level results can then be aggregated to give the figures below.

We distinguish the percentage of the workforce in occupations predicted to see a rise in workforce share in 2030 with a 'low probability' (less than 30%), 'medium probability' ($> 50\%$) and 'high probability' ($> 70\%$). 30% and 70% are also the thresholds used in (Frey and M. A. Osborne 2017). That is, we calculate total employment that has probability of future increasing demand lying above and below these three thresholds.

As per Section 6.3.2, our calculations are made at the finest level available for US occupations; that is, the six-digit US SOC 2010. The percentage of the US workforce as partitioned by the thresholds above is provided in Table 6.2.

Table 6.2: The fraction of the US workforce above and below varying thresholds for the probability of increasing demand.

Number of Occupations	Employment	below 0.3	above 0.5	above 0.7
772	135 million	18.7%	43.2%	9.6%

In Figure 6.4, we plot (following (Frey and M. A. Osborne 2017)) the distribution of current US employment over its probability of future increased demand. We

additionally distinguish this employment by an intermediate aggregation of the Major Groups, as specified by BLS 2010 Standard Occupational Classification user guide (U.S. Bureau of Labor Statistics 2010).

Figure 6.4: The distribution of US employment according to its probability of future increased demand. Note that the total area under all curves is equal to total US employment. This figure illustrates a high degree of uncertainty about occupational demand captured by our approach with a large proportion of employment mass centred around 0.5. This contrasts sharply with a much more certain 'either or' U-shaped distribution by probability of automation in (Frey and M. A. Osborne 2017).

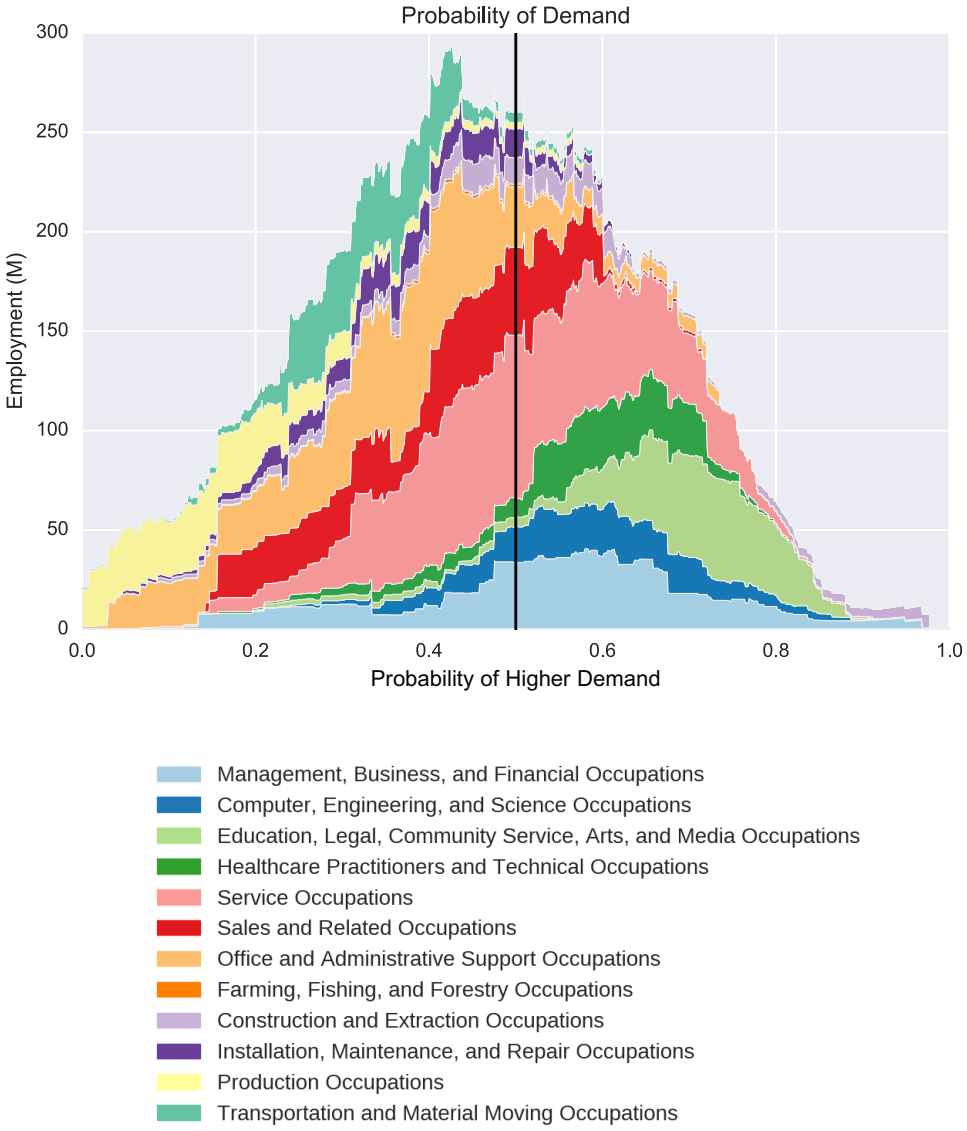


Figure 6.4 reveals a large mass of the workforce in employment with highly

uncertain demand prospects (that is, a probability of experiencing higher workforce share of around 0.5). Note that this contrasts sharply with the U-shaped distribution by probability of automation in (Frey and M. A. Osborne 2017) and (Frey and M. A. Osborne 2014), where the workforce is overwhelmingly in occupations at either a very high probability or a very low probability of automation. That our predictions are more uncertain is a direct result of the distinctions of our methodology from previous work. Firstly, the expert labels we gather in our foresight exercises (see Section 6.3.3) are explicitly clothed in uncertainty, whereas (Frey and M. A. Osborne 2017) and (Frey and M. A. Osborne 2014) assume that participants are certain about their labels. This humility is partially motivated by the difficulty of the task assigned to our experts: balancing all the macro trends that might influence the future of work. Our allowance for our experts' self-assessed degrees of confidence also recognises that many of the macro trends act at cross-purposes, leading to uncertainty about which will dominate in the case of any one occupation. Secondly, we use 120 O*NET features, against the nine used in (Frey and M. A. Osborne 2017) and (Frey and M. A. Osborne 2014). This more detailed characterisation of occupations renders occupations less similar to one another, and hence limits the confidence of our model in predicting for one occupation based on what has been labelled for another.

Table 6.3 lists the minor occupation groups about which our model is most optimistic.

Table 6.3: For the US, the minor occupation groups with the greatest probabilities of future increased demand. For these occupations, we characterise the fraction of their current employment that has a probability of increased demand above two thresholds.

Title	Employment	> 0.7	> 0.5
Preschool, Primary, Secondary, And Special Education School Teachers	4,050,880	97.8	100
Animal Care And Service Workers	185,780	93.7	100
Lawyers, Judges, And Related Workers	672,580	90.7	98.1
Postsecondary Teachers	1,328,890	83.0	100
Engineers	1,610,470	70.0	100
Personal Appearance Workers	504,640	69.0	100

Social Scientists And Related Workers	239,170	65.6	92
Counselors, Social Workers, And Other Community And Social Service Specialists	1,715,190	54.0	100
Librarians, Curators, And Archivists	253,800	51.8	62.9
Entertainers And Performers, Sports And Related Workers	483,450	46.4	96.1
Other Management Occupations	2,185,950	42.9	100
Media And Communication Workers	542,570	40.3	89.4
Operations Specialties Managers	1,663,790	29.8	46.5
Religious Workers	, 68,530	29.6	100
Other Teachers And Instructors	282,640	23.0	100
Other Personal Care And Service Workers	2,619,120	21.9	100
Construction Trades Workers	4,076,790	21.8	64.7
Business Operations Specialists	4,424,800	19.6	77.4
Physical Scientists	266,050	13.8	100
Other Sales And Related Workers	585,030	12.3	14.4
Architects, Surveyors, And Cartographers	168,650	11.8	67.3
Other Education, Training, And Library Occupations	1,386,830	10.1	100
Other Healthcare Support Occupations	1,451,710	6.3	54.3
Occupational Therapy And Physical Therapist Assistants And Aides	174,800	4.3	100
Health Diagnosing And Treating Practitioners	4,944,470	4.0	100

We derive a number of insights from Table 6.3, informed in part by our workshop discussions.

- Education and personal care occupations feature prominently in the rankings; however, healthcare occupations are lower than expected by trends such as ageing, potentially reflecting uncertainty over the trajectory of healthcare policy and spending in the US or technical issues related to the composition of the training set (which in practice under-represented healthcare occupations).

- Construction trade work, as a larger employer, is another beneficiary. It is supported by a number of trends, including urbanisation, ageing and globalisation and is expected to be an important source of medium-skilled jobs in the future.

- Demand prospects can vary considerably for occupations which are otherwise very similar. For example, business operations specialists – which typically need information management expertise – are set to grow as a share of the workforce while neighbouring minor occupation groups in the SOC such as financial specialists (see Table 6.4) are predicted to fall in share. Looking at the detailed occupation level, the results for business operations specialists are driven by management analysts, training and development specialists, labor relations specialists, logisticians and meeting, convention and event planners in particular – occupations which will conceivably benefit from the reorganisation of work and the workplace.

- Another niche anticipated to grow in workforce share is other sales and related workers, and within that in particular sales engineers and real estate agents, notwithstanding the predicted decline in general sales occupations.

Table 6.4: For the US, the minor occupation groups with the lowest probabilities of future increased demand. We characterise for these occupations the fraction of their current employment that has a probability of increased demand below two thresholds.

Title	Employment	< 0.3	< 0.5
Woodworkers	236,460	100%	100%
Printing Workers	256,040	100%	100%
Metal Workers And Plastic Workers	1,923,050	98.7%	100%
Financial Clerks	3,144,540	97.7%	100%
Other Production Occupations	2,552,400	96.9%	99.4%
Plant And System Operators	311,060	94.1%	100%
Assemblers And Fabricators	1,571,480	92.2%	100%
Communications Equipment Operators	110,250	91.2%	100%
Food Processing Workers	738,030	89.1%	100%
Forest, Conservation, And Logging Workers,	42,740	83.9%	100%
Textile, Apparel, And Furnishings Workers	561,550	81.5%	100%
Extraction Workers	253,530	66.7%	100%
Financial Specialists	2,607,770	66.3%	90.7%
Rail Transportation Workers	117,460	53.2%	100%
Cooks And Food Preparation Workers	3,132,040	49.7%	100%
Sales Representatives, Services	1,808,330	49.0%	100%
Other Transportation Workers	305,320	48.3%	100%
Retail Sales Workers	8,799,240	44.9%	47.6%
Other Construction And Related Workers	393,710	39.8%	63.2%
Water Transportation Workers	77,270	39.6%	100%
Vehicle And Mobile Equipment Mechanics, Installers, And Repairers	1,554,340	38.0%	99.2%
Librarians, Curators, And Archivists	253,800	37.1%	37.1%
Material Recording, Scheduling, Dispatching, And Distributing Workers	3,973,730	32.1%	97.6%
Other Installation, Maintenance, And Repair Occupations	2,776,890	28.4%	90%
Entertainment Attendants And Related Workers	524,310	25.2%	96.7%
Motor Vehicle Operators	3,797,540	24.3%	100%
Material Moving Workers	4,473,640	20.9%	100%
Other Office And Administrative Support Workers	3,723,230	20.2%	100%
Agricultural Workers	383,890	17.9%	100%
Construction Trades Workers	4,076,790	8.8%	35.3%

Other Healthcare Support Occupations	1,451,710	7.5%	45.7%
Health Technologists And Technicians	2,909,230	6.5%	56.3%
Information And Record Clerks	5,336,050	6.4%	95%
Secretaries And Administrative Assistants	3,680,630	5.5%	100%
Legal Support Workers	344,220	5.1%	100%
Electrical And Electronic Equipment Mechanics, Installers, And Repairers	585,280	3.1%	100%
Business Operations Specialists	4,424,800	2.9%	22.6%
Other Protective Service Workers	1,524,350	2.7%	89.8%
Grounds Maintenance Workers	959,960	2.5%	6.7%
Drafters, Engineering Technicians, And Mapping Technicians	680,790	2.2%	74.3%
Life, Physical, And Social Science Technicians	359,460	1.8%	82.7%
Other Transportation Workers	305,320	1%	100%

- These results support the importance of future routine-biased technological change. Notably they anticipate the impact of automation encroaching on more cognitively advanced and complex occupations such as financial specialists.

- The predicted fall in retail sales workers and entertainment attendants, which between them account for a large volume of employment, is consistent with an expansion in digitally provided goods and services.

- The transportation occupations represented may reflect a belief that driverless cars will disrupt the future workforce. The rise of the sharing economy might reasonably be expected to lead to an increased demand for installation and reparation jobs, especially in areas like transport as cars and other assets are used more intensively, but this hypothesis is not supported here.

6.5.2 Sensitivity analysis

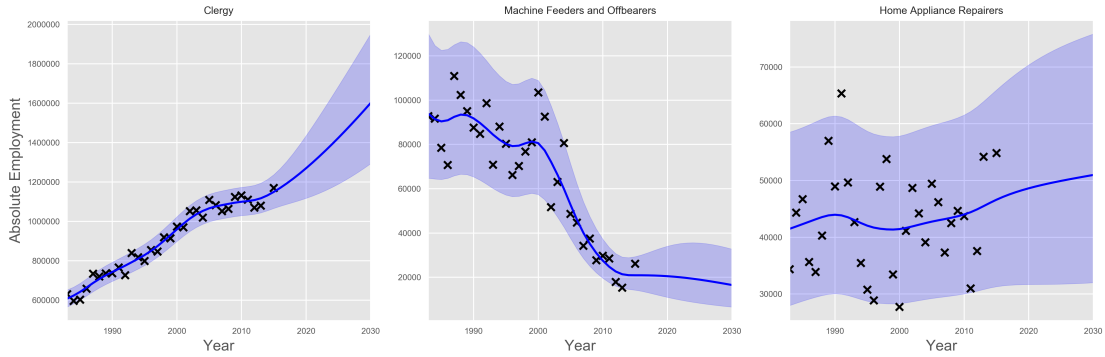
Firstly, to test for the generalisability of our results, we perform a cross-validation exercise. In particular, we randomly select a reduced training set of half the available data (corresponding to the labels for fifteen occupations); the remaining data form a test set. On the test set, we evaluate the receiver operating characteristic (ROC) curve (Murphy 2012). Given that we observe ternary, as opposed to binary, labels,

the receiver operating characteristic (ROC) curve is a surface (Waegeman et al. 2008). An approximation is made to the Volume Under the Surface (VUS), whereby the Area Under the Curve (AUC) is calculated separately for each ternary component. The volume under the ROC surface is taken to be the mean of the separate Area Under the Curve (AUC) slices. As the workshop data or model prediction for an occupation is either an empirical or posterior distribution over ternary labels respectively, samples are drawn and the mean Volume Under the Surface (VUS) is calculated. Due to uncertainty in the distribution, a normalised VUS is calculated by dividing the test VUS by the ground truth VUS. We repeat this experiment for fifty random splits of the thirty occupations from the workshop set. Each experiment is composed of ten training occupations and twenty test occupations. A high AUC/VUS (which ranges from 0.5 to 1) indicates that our model is able to reliably predict fifteen occupations given a distinct set of fifteen occupations. This would suggest both that the model is effective and that the training set is self-consistent. It would also imply a degree of robustness of our results to the inclusion (or exclusion) of a small number of occupations in that training set.

We also investigate how breaks in long-run trends as perceived by the workshop experts contribute to the findings above. We do this by re-running the predictive model but using non-parametric extrapolation (as described in Section 6.4.5) of occupational employment to label the occupations in place of the experts' judgments.

As an example of extrapolated trends, Figure 6.5 shows the absolute employment extrapolations for three US occupations, one each for higher, same and lower probability. The shaded areas give the 90% credibility interval around the mean.

Figure 6.5: Three US occupations for absolute employment number extrapolations. We show an occupation with probability of higher, same and lower in 2030. The shaded area is the 95% confidence interval. - Created by Justin Bewshew



Note that, in what follows, all rankings of occupation groups from our model use the employment-weighted average of probabilities of occupations within the group.

Firstly, Table 6.5 presents the results of our cross-validation exercise on US workshop data. The high VUS results suggest that our model is accurate, and that our results are not sensitive to small changes to the training set.

Table 6.5: The volume under the receiver operating characteristic (ROC) from cross-validation of model on US workshop data.

Mean	Lower Quartile	Upper Quartile
0.949	0.943	0.956

Table 6.6 compares the probability of increased demand generated by trend extrapolation against that obtained from the judgment of experts.

Table 6.6: The percentage point differences (of trend extrapolation from our workshop-trained predictions) in the probabilities of future demand at minor occupation group level. That is, each row is the probability produced by trend extrapolation minus the probability produced by the workshop, multiplied by 100.

Title	pp difference in probability of demand
Other Office And Administrative Support Workers	-52.2
Legal Support Workers	-47.8
Financial Clerks	-47.0

Table 6.6: The percentage point differences (of trend extrapolation from our workshop-trained predictions) in the probabilities of future demand at minor occupation group level. That is, each row is the probability produced by trend extrapolation minus the probability produced by the workshop, multiplied by 100.

Title	pp difference in probability of demand
Communications Equipment Operators	-42.2
Financial Specialists	-39.6
Information And Record Clerks	-39.2
Secretaries And Administrative Assistants	-38.7
Sales Representatives, Services	-35.6
Entertainment Attendants And Related Workers	-34.4
Retail Sales Workers	-32.7
Rail Transportation Workers	-27.7
Motor Vehicle Operators	-27.5
Other Protective Service Workers	-27.4
Material Recording, Scheduling, Dispatching, And Distributing Workers	-27.4
Sales Representatives, Wholesale And Manufacturing	-26.3
Librarians, Curators, And Archivists	-26.2
Supervisors Of Transportation And Material Moving Workers	-26.1
Other Healthcare Support Occupations	-25.7
Other Transportation Workers	-25.4
Assemblers And Fabricators	-24.9
Other Production Occupations	-23.6
Lawyers, Judges, And Related Workers	-23.1
Baggage Porters, Bellhops, And Concierges	-22.8
Food And Beverage Serving Workers	-22.8
Supervisors Of Office And Administrative Support Workers	-22.4
Printing Workers	-22.1
Forest, Conservation, And Logging Workers	-21.6
Health Technologists And Technicians	-21.2
Food Processing Workers	-21.1
Supervisors Of Sales Workers	-20.7
Operations Specialties Managers	-20.4
Supervisors Of Personal Care And Service Workers	-20.3
Cooks And Food Preparation Workers	-20.2

Table 6.6: The percentage point differences (of trend extrapolation from our workshop-trained predictions) in the probabilities of future demand at minor occupation group level. That is, each row is the probability produced by trend extrapolation minus the probability produced by the workshop, multiplied by 100.

Title	pp difference in probability of demand
Top Executives	-19.7
Supervisors Of Food Preparation And Serving Workers	-19.6
Religious Workers	-19.3
Other Sales And Related Workers	-19.2
Other Construction And Related Workers	-19.2
Media And Communication Workers	-18.9
Plant And System Operators	-18.7
Law Enforcement Workers	-18.5
Business Operations Specialists	-18.1
Textile, Apparel, And Furnishings Workers	-18.1
Counselors, Social Workers, And Other Community And Social Service Specialists	-18.0
Extraction Workers	-17.9
Material Moving Workers	-17.2
Other Management Occupations	-17.1
Metal Workers And Plastic Workers	-17.1
Supervisors Of Production Workers	-16.8
Water Transportation Workers	-16.3
Fishing And Hunting Workers	-15.9
Tour And Travel Guides	-15.8
Nursing, Psychiatric, And Home Health Aides	-15.6
Other Teachers And Instructors	-15.4
Health Diagnosing And Treating Practitioners	-15.4
Social Scientists And Related Workers	-15.2
Supervisors Of Protective Service Workers	-15.2
Woodworkers	-15.1
Vehicle And Mobile Equipment Mechanics, Installers, And Repairers	-14.8
Media And Communication Equipment Workers	-14.5
Agricultural Workers	-14.4
Other Food Preparation And Serving Related Workers	-14.4
Supervisors Of Construction And Extraction Workers	-14.2

Table 6.6: The percentage point differences (of trend extrapolation from our workshop-trained predictions) in the probabilities of future demand at minor occupation group level. That is, each row is the probability produced by trend extrapolation minus the probability produced by the workshop, multiplied by 100.

Title	pp difference in probability of demand
Occupational Therapy And Physical Therapist Assistants And Aides	-14.1
Supervisors Of Building And Grounds Cleaning And Maintenance Workers	-13.9
Advertising, Marketing, Promotions, Public Relations, And Sales Managers	-13.8
Other Education, Training, And Library Occupations	-13.7
Life, Physical, And Social Science Technicians	-13.6
Supervisors Of Farming, Fishing, And Forestry Workers	-13.5
Air Transportation Workers	-13.4
Other Healthcare Practitioners And Technical Occupations	-13.3
Preschool, Primary, Secondary, And Special Education School Teachers	-12.2
Other Personal Care And Service Workers	-11.9
Fire Fighting And Prevention Workers	-10.9
Entertainers And Performers, Sports And Related Workers	-10.8
Funeral Service Workers	-10.6
Other Installation, Maintenance, And Repair Occupations	-10.4
Construction Trades Workers	-9.9
Drafters, Engineering Technicians, And Mapping Technicians	-8.4
Electrical And Electronic Equipment Mechanics, Installers, And Repairers	-7.8
Life Scientists	-7.3
Art And Design Workers	-6.8
Helpers, Construction Trades	-6.5
Supervisors Of Installation, Maintenance, And Repair Workers	-5.9
Architects, Surveyors, And Cartographers	-5.8
Mathematical Science Occupations	-4.5
Computer Occupations	-4.2
Grounds Maintenance Workers	-3.5

Table 6.6: The percentage point differences (of trend extrapolation from our workshop-trained predictions) in the probabilities of future demand at minor occupation group level. That is, each row is the probability produced by trend extrapolation minus the probability produced by the workshop, multiplied by 100.

Title	pp difference in probability of demand
Postsecondary Teachers	-3.3
Physical Scientists	-1.9
Personal Appearance Workers	-1.0
Building Cleaning And Pest Control Workers	4.2
Animal Care And Service Workers	4.4
Engineers	8.8

Table 6.7 further ranks occupation groups, using the same intermediate aggregation of major groups as described in Section 6.5.1, by their probability of rising demand using our expert judgment training set, and compares these against rankings based on independent quantitative forecasts for the year 2024 from the US Bureau of Labor Statistics (BLS).

Table 6.7: Relative rankings of intermediate aggregation of major occupation groups in the US by our model, trained on expert judgment, and by independent forecasts from the BLS.

Ranking from Expert Judgment	Ranking from BLS Projections 2014-2024
Education, Legal, Community Service, Arts, and Media Occupations	Healthcare practitioners and technical occupations
Computer, Engineering, and Science Occupations	Construction and extraction occupations
Healthcare Practitioners and Technical Occupations	Service Occupations
Management, Business, and Financial Occupations	Computer, Engineering and Service Occupations
Construction and Extraction Occupations	Education, Legal, Community Service, Arts, and Media Occupations
Service Occupations	Management, Business and Financial operations
Sales and Related Occupations	Installation, maintenance, and repair occupations
Farming, Fishing, and Forestry Occupations	Sales and related occupations
Installation, Maintenance, and Repair Occupations	Transportation and material moving occupations

Table 6.7: Relative rankings of intermediate aggregation of major occupation groups in the US by our model, trained on expert judgment, and by independent forecasts from the BLS.

Ranking from Expert Judgment	Ranking from BLS Projections 2014-2024
Office and Administrative Support Occupations	Office and administrative support occupations
Transportation and Material Moving Occupations	Production occupations
Production Occupations	Farming, fishing, and forestry occupations

Tables 6.6 and 6.7 reveal the following insights.

- Firstly, Table 6.6 makes it clear that expert judgment is considerably more pessimistic than trend extrapolation for most minor occupation groups. The divergence is largest for routine cognitive, as opposed to routine manual, occupations.

- We see considerable divergence across the three outlooks (our workshop-informed model, our trend extrapolation, and the BLS projection). This is arguably most marked for occupations in *legal*, *architecture and engineering* and *arts, design, entertainment, and sports and media*. There is seemingly greater agreement, however, over occupations projected to decline in workforce share.

6.5.3 Skills

We now describe the findings of our study on the relationships between O*NET variables (which we refer to as 'features' and occasionally informally refer to as 'skills') and future demand. Note that our methodology (see Section 6.4.3) provides employment-weighted results: as such, all the results which follow are robust to outlying occupations with small employment.

We use two measures of the importance of features to future demand: the Pearson correlation coefficient (Section 6.4.3) and the average derivative (Section 6.4.3). In interpreting the values of these measures, note firstly that the correlation

coefficient lies between -1 and 1. The average derivative is calculated by considering the derivative of an unobservable real-valued function linked to demand. It is dimensionless; an average derivative's magnitude is significant only relative to that of another average derivative. For either measure, positive values are associated with features whose increase is expected to increase demand, and negative values with features whose increase is expected to decrease demand.

As described in Section 6.4.3, we exclude especially noisy features from consideration under the average derivative measure.

These variables are:

- Perceptual Speed (Abilities);
- Building and Construction (Knowledge);
- Food Production (Knowledge);
- Production and Processing (Knowledge);
- Control Precision (Abilities);
- Biology (Knowledge), and
- Fine Arts (Knowledge).

Note that the excluded variables are predominately knowledge features. The significance of knowledge features, perhaps more than other, differs considerably across occupations. It is hence not surprising that these features are more likely to be less reliable than others under the average derivative metric.

US

Table 6.8: A ranking, by Pearson correlation, of the importance of O*NET variables to future demand for US occupations.

Rank	O*NET Variable	Class	Pearson Correlation
1	Learning Strategies	Skills	0.632
2	Psychology	Knowledge	0.613

Table 6.8: A ranking, by Pearson correlation, of the importance of O*NET variables to future demand for US occupations.

Rank	O*NET Variable	Class	Pearson Correlation
3	Instructing	Skills	0.609
4	Social Perceptiveness	Skills	0.605
5	Sociology and Anthropology	Knowledge	0.603
6	Education and Training	Knowledge	0.602
7	Coordination	Skills	0.571
8	Originality	Abilities	0.570
9	Fluency of Ideas	Abilities	0.562
10	Active Learning	Skills	0.534
11	Therapy and Counseling	Knowledge	0.531
12	Philosophy and Theology	Knowledge	0.526
13	Speaking	Skills	0.514
14	Service Orientation	Skills	0.511
15	Active Listening	Skills	0.507
16	Complex Problem Solving	Skills	0.502
17	Oral Expression	Abilities	0.493
18	Communications and Media	Knowledge	0.491
19	Speech Clarity	Abilities	0.489
20	Judgment and Decision Making	Skills	0.482
21	English Language	Knowledge	0.474
22	Monitoring	Skills	0.470
23	Deductive Reasoning	Abilities	0.468
24	Oral Comprehension	Abilities	0.465
25	Critical Thinking	Skills	0.462
26	Systems Evaluation	Skills	0.461
27	History and Archeology	Knowledge	0.452
28	Inductive Reasoning	Abilities	0.448
29	Persuasion	Skills	0.443
30	Speech Recognition	Abilities	0.436
31	Science	Skills	0.431
32	Negotiation	Skills	0.419
33	Management of Personnel Resources	Skills	0.418
34	Systems Analysis	Skills	0.415
35	Problem Sensitivity	Abilities	0.414
36	Writing	Skills	0.407
37	Operations Analysis	Skills	0.395

Table 6.8: A ranking, by Pearson correlation, of the importance of O*NET variables to future demand for US occupations.

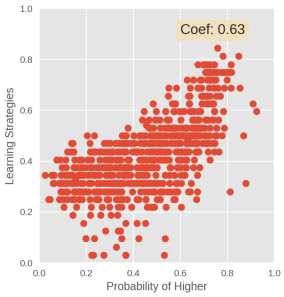
Rank	O*NET Variable	Class	Pearson Correlation
38	Administration and Management	Knowledge	0.388
39	Biology	Knowledge	0.388
40	Fine Arts	Knowledge	0.385
41	Reading Comprehension	Skills	0.374
42	Memorization	Abilities	0.372
43	Time Management	Skills	0.360
44	Foreign Language	Knowledge	0.359
45	Written Expression	Abilities	0.351
46	Medicine and Dentistry	Knowledge	0.348
47	Technology Design	Skills	0.345
48	Personnel and Human Resources	Knowledge	0.344
49	Written Comprehension	Abilities	0.341
50	Information Ordering	Abilities	0.328
51	Time Sharing	Abilities	0.316
52	Geography	Knowledge	0.310
53	Law and Government	Knowledge	0.309
54	Customer and Personal Service	Knowledge	0.291
55	Category Flexibility	Abilities	0.284
56	Speed of Closure	Abilities	0.268
57	Management of Material Resources	Skills	0.262
58	Chemistry	Knowledge	0.192
59	Public Safety and Security	Knowledge	0.189
60	Telecommunications	Knowledge	0.189
61	Computers and Electronics	Knowledge	0.186
62	Management of Financial Resources	Skills	0.160
63	Design	Knowledge	0.146
64	Flexibility of Closure	Abilities	0.133
65	Physics	Knowledge	0.126
66	Programming	Skills	0.122
67	Engineering and Technology	Knowledge	0.121
68	Visualization	Abilities	0.120
69	Sales and Marketing	Knowledge	0.118
70	Far Vision	Abilities	0.105
71	Explosive Strength	Abilities	0.099
72	Building and Construction	Knowledge	0.078

Table 6.8: A ranking, by Pearson correlation, of the importance of O*NET variables to future demand for US occupations.

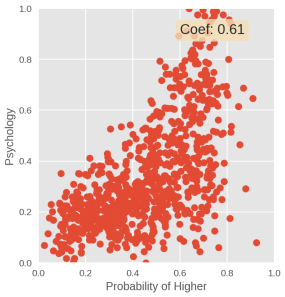
Rank	O*NET Variable	Class	Pearson Correlation
73	Selective Attention	Abilities	0.069
74	Clerical	Knowledge	0.047
75	Auditory Attention	Abilities	0.036
76	Economics and Accounting	Knowledge	0.036
77	Mathematical Reasoning	Abilities	0.035
78	Near Vision	Abilities	0.016
79	Mathematics – Skills	Skills	0.008
80	Transportation	Knowledge	0.004
81	Mathematics – Knowledge	Knowledge	-0.006
82	Number Facility	Abilities	-0.022
83	Dynamic Flexibility	Abilities	-0.023
84	Quality Control Analysis	Skills	-0.028
85	Stamina	Abilities	-0.033
86	Food Production	Knowledge	-0.034
87	Trunk Strength	Abilities	-0.039
88	Gross Body Coordination	Abilities	-0.059
89	Gross Body Equilibrium	Abilities	-0.063
90	Visual Color Discrimination	Abilities	-0.081
91	Installation	Skills	-0.082
92	Dynamic Strength	Abilities	-0.111
93	Troubleshooting	Skills	-0.114
94	Extent Flexibility	Abilities	-0.129
95	Equipment Selection	Skills	-0.141
96	Static Strength	Abilities	-0.142
97	Hearing Sensitivity	Abilities	-0.142
98	Mechanical	Knowledge	-0.152
99	Perceptual Speed	Abilities	-0.168
100	Depth Perception	Abilities	-0.173
101	Speed of Limb Movement	Abilities	-0.185
102	Spatial Orientation	Abilities	-0.198
103	Sound Localization	Abilities	-0.207
104	Multilimb Coordination	Abilities	-0.219
105	Production and Processing	Knowledge	-0.239
106	Operation Monitoring	Skills	-0.242
107	Night Vision	Abilities	-0.244

Table 6.8: A ranking, by Pearson correlation, of the importance of O*NET variables to future demand for US occupations.

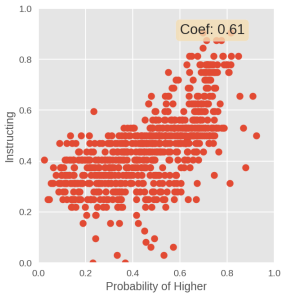
Rank	O*NET Variable	Class	Pearson Correlation
108	Peripheral Vision	Abilities	-0.246
109	Glare Sensitivity	Abilities	-0.247
110	Repairing	Skills	-0.259
111	Response Orientation	Abilities	-0.282
112	Equipment Maintenance	Skills	-0.284
113	Arm-Hand Steadiness	Abilities	-0.297
114	Reaction Time	Abilities	-0.322
115	Operation and Control	Skills	-0.326
116	Finger Dexterity	Abilities	-0.354
117	Manual Dexterity	Abilities	-0.365
118	Rate Control	Abilities	-0.394
119	Wrist-Finger Speed	Abilities	-0.423
120	Control Precision	Abilities	-0.466



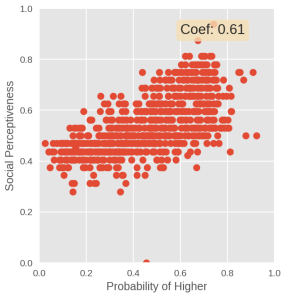
1. Learning Strategies



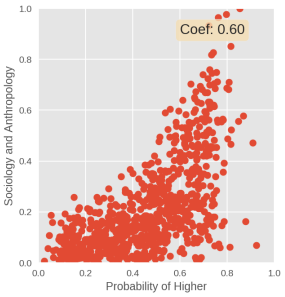
2. Psychology



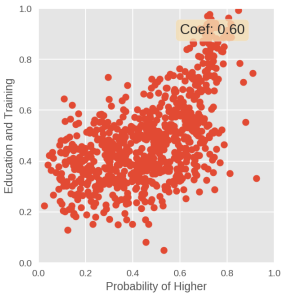
3. Instructing



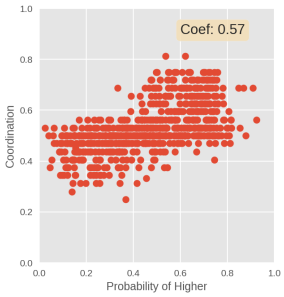
4. Social Perceptiveness



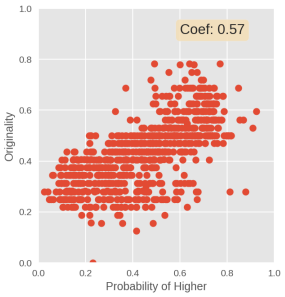
5. Sociology and Anthropology



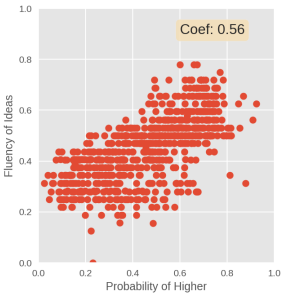
6. Education and Training



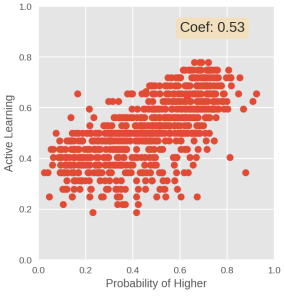
7. Coordination



8. Originality

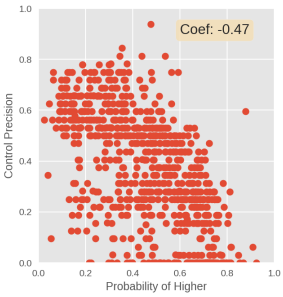


9. Fluency of Ideas

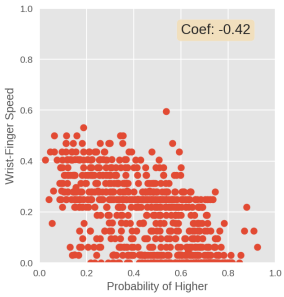


10. Active Learning

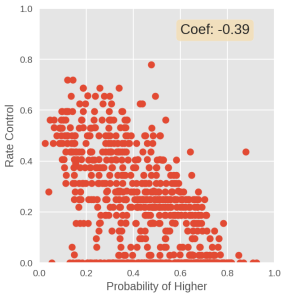
Figure 6.6: The ten most important O*NET variables as ranked by Pearson correlation for the US.



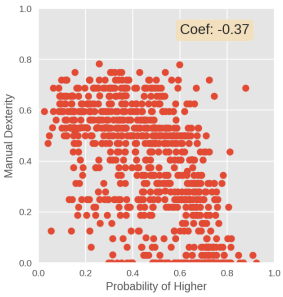
1. Control Precision



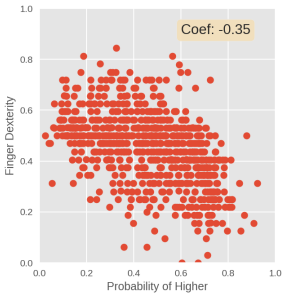
2. Wrist Finger Speed



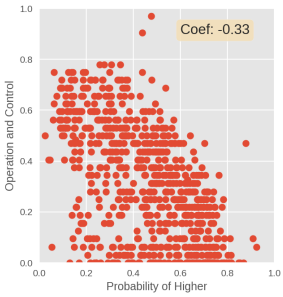
3. Rate Control



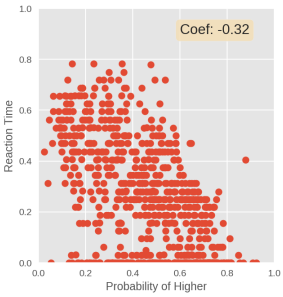
4. Manual Dexterity



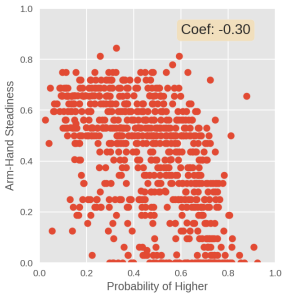
5. Finger Dexterity



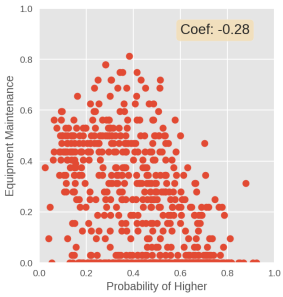
6. Operation and Control



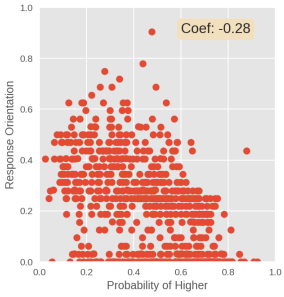
7. Reaction Time



8. Arm Hand Steadiness



9. Equipment Maintenance



10. Response Orientation

Figure 6.7: The ten least important O*NET variables as ranked by Pearson correlation for the US.

Table 6.8 ranks, by Pearson correlation coefficient, all 120 variables according to their association with a rising occupation workforce share (in declining order of strength). Figures 6.6 and 6.7 plot, respectively, the top and bottom ten O*NET variables as ranked by Pearson correlation. Table B.1 in the Appendix also provides aggregate rankings for the average derivative.

- The results confirm the importance of 21st century skills in the US, with a particularly strong emphasis on interpersonal competencies. This is underscored by the presence of skills such as Instructing, Social Perceptiveness and Coordination, and related knowledge domains such as Psychology, Sociology and Anthropology.

- This is consistent with the literature on the increasing importance of social skills – recall, the fact that between 1980 and 2012, jobs with high social skill requirements grew by nearly 10 percentage points as a share of the US labour force (Deming 2015). There are good reasons to think that these trends will continue – not only as organisations seek to reduce the costs of coordination but also as they negotiate the cultural context in which globalisation and the spread of digital technology are taking shape (Tett 2017). A variety of interventions targeted at different stages of the life cycle have proven successful in fostering social skills. The evidence base is largest on the success of early programmes. Workplace-based internships and apprenticeships, however, also have a good track record arising from the need to learn informal or tacit knowledge and the bonds of attachment between a supervisor and an apprentice (Kautz et al. 2014).

- The results also emphasise the importance of higher-order cognitive skills such as Originality and Fluency of Ideas. Learning Strategies and Active Learning – the ability of students to set goals, ask relevant questions, get feedback as they learn and apply that knowledge meaningfully in a variety of contexts- also feature prominently.

- Progress towards developing these skills as part of the formal education system has been slow due to difficulties in understanding how they arise and develop over time and how they can be embedded in the curriculum and formal assessments. Nonetheless, a number of initiatives have shown promise and are beginning to shape domestic and international policy dialogue (Schunk and Zimmerman 2007; Lucas et al. 2013; OECD 2016a). Strengthening the affective aspects of education and a lifelong learning habit, especially among boys and students from disadvantaged backgrounds who tend to have lower levels of motivation, is a further area of interest for policymakers. The research literature shows that teachers can play an important role – both in raising student expectations and in rewarding the process of learning – for instance, in giving students opportunities to share the results of their work with others or explain why what they learned was valuable to them, though they are unlikely to be sufficient in the absence of other policies to promote educational excellence and equity (Covington and Müeller 2001; Diamond et al. 2004; Weinstein 2002; Hampden-Thompson and Bennett 2013; OECD 2017).

- In addition to knowledge fields related to social skills, English language, History and Archeology, Administration and Management and Biology are all associated strongly with occupations predicted to see a rise in workforce share, reminding us that the future workforce will have generic knowledge as well as skills requirements.

- Psychomotor and physical abilities are strongly associated with occupations with a falling workforce share. Interestingly, this includes abilities such as Finger Dexterity and Manual Dexterity, which (Frey and M. A. Osborne 2017) identified as key bottlenecks to automation. Trade and offshoring offer a potential explanation for why these skills might fall in demand - consistent with workshop participants having considered a broad range of trends. The main feature that makes a job potentially offshorable or vulnerable to import competition hinges less on a task's routineness or non-routineness than the cost

advantages of producing overseas and the marginal importance of face-to-face interactions in the production process.

- The correlations for variables associated with a rising occupation workforce share are in general stronger than those associated with a falling occupation workforce share. This is perhaps not surprising: *ceteris paribus*, an increase in the value of any O*NET variable for an occupation makes it more skilled, and might broadly be expected to result in greater demand (even if there are other reasons why the occupation will experience a fall in demand). It is also fortunate: our core emphasis is on informing skills policy, which has a natural focus on those skills most strongly linked to growing demand.

6.5.4 Relative importance of knowledge, skills and abilities

We now provide a comparison of the overall relative importance of knowledge, skills and abilities as captured in O*NET. All figures feature on the horizontal axis the rank of a feature: the further to the right, the less important it is for demand. This importance is assessed using linear (Pearson correlation coefficient) and non-linear (average derivative) metrics. The vertical axis shows the fractions of features in a sliding window (over rankings) that are, respectively, knowledge, skills and abilities.

In Figures 6.8 and 6.9, it can be seen that abilities are broadly less important (weighted to the right). Perhaps the most interesting insight, however, is that the non-linear metric (the average derivative) gives knowledge features more weight to the left. That is, a non-linear measure ranks knowledge features more highly. Recall that the benefit of a non-linear metric is that it allows us to discover complementarities: skills that are only important if other skills take high values. As such, this result is compatible with the intuition that knowledge features (like Psychology and Foreign Language) are mostly valuable as complements. We find a similar pattern in a large number of STEM-related features (like Science, Technology Design and Operations Analysis). They are not equally useful to all occupations (as would be required to be assessed as important for our linear metric), but find use only for some specialised occupations that have high values for other skills.

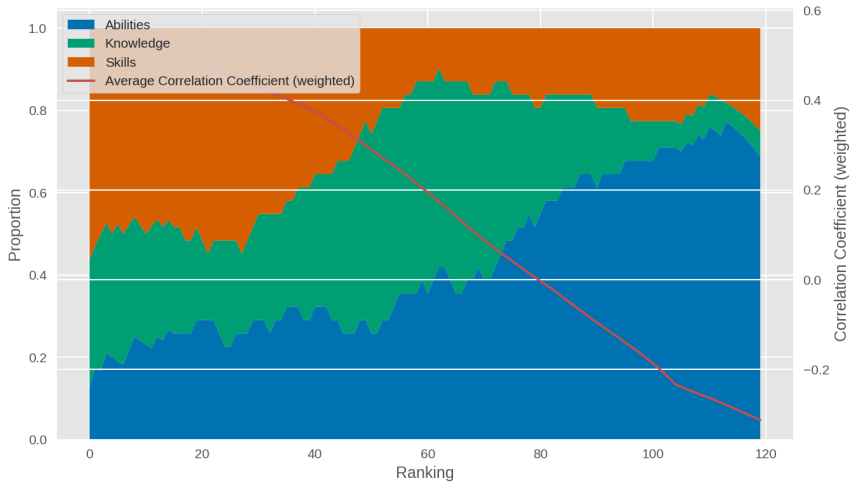


Figure 6.8: The relative importance of knowledge, skills and abilities as assessed by Pearson correlation coefficient for the US.

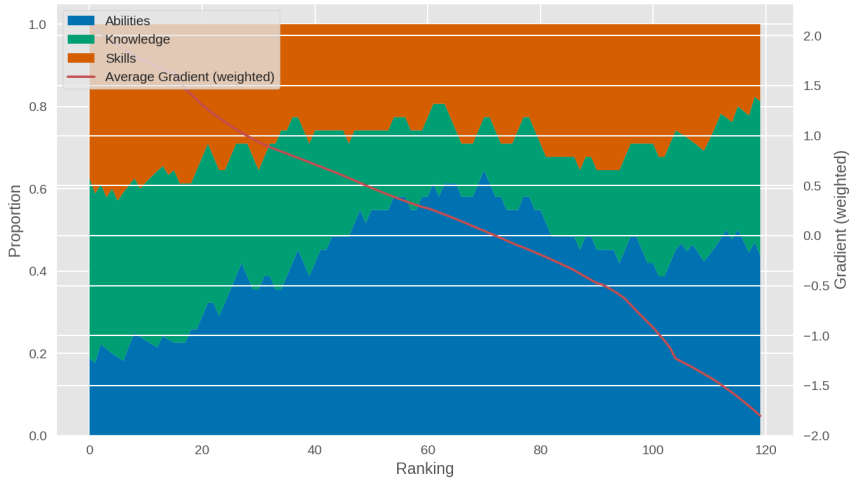


Figure 6.9: The relative importance of knowledge, skills and abilities as assessed by average derivative for the US.

6.5.5 Skill complementarities

Recall from Section 6.4.3, that we say that an O*NET feature a is complementary to an O*NET feature b if increasing a increases demand for occupations with large values of b. Conversely, a is anti-complementary to an O*NET feature b if increasing a decreases demand for occupations with large values of b. For each

sub-major occupation group, we rank the three features that would most drive a: (i) rising workforce share for a unit increase in the feature (complementary features), and (ii) falling workforce share for a unit increase in the feature (anti-complementary features). As such, we are also able to establish which features are most important for different regions of the skill space. We represent the position of an occupation group in skills space by listing its highest ranked features, which we term its *current features*.

US

We describe in Table 6.9 complementary and anti-complementary O*NET variables for US sub-major occupation groups.

Take Production Occupations, for example, which Figure 6.4 shows are predicted to see a fall in workforce share. According to the O*NET data, Production and Processing, Near Vision and Problem Sensitivity are the three most important or emblematic features for this occupation group. Our model predicts that increasing Customer and Personal Service, Technology Design and Installation (Rate Control, Operation and Control and Quality Control Analysis) in the presence of these features will have the greatest positive (or negative) impact on future demand. In fact, looking across all occupation groups, Customer and Personal Service and Technology Design (along with Science) appear to be the O*NET features most likely to appear as positive complementary variables.

Of course, any reconfiguration of skills and knowledge requirements entails an evolution of the occupation. Or put differently, occupations may need to be redesigned in order to make effective use of skills and knowledge complements – and the results presented in Table 6.9 could be a useful guide in this exercise.

Table 6.9: For US sub-major occupation groups, ranked lists (the highest ranked, top, and lowest ranked, bottom) of O*NET features that are: currently high-valued; complementary and; anti-complementary.

SOC	Title	Current features	Complementary features	Anti-complementary features
27-0000	Arts, Design, Entertainment, Sports, And Media Occupations	English Language	Science	Economics and Accounting
		Oral Expression	Philosophy and Theology	Rate Control
		Oral Comprehension	Education and Training	Mathematics – Knowledge
29-0000	Healthcare Practitioners And Technical Occupations	Medicine and Dentistry	Technology Design	Medicine and Dentistry
		Customer and Personal Service	Science	Rate Control
		Oral Comprehension	Operations Analysis	Operation and Control
13-0000	Business And Financial Operations Occupations	Oral Comprehension	Science	Medicine and Dentistry
		Written Comprehension	Philosophy and Theology	Economics and Accounting
		English Language	Technology Design	Mathematics – Knowledge
53-0000	Transportation And Material Moving Occupations	Multilimb Coordination	Customer and Personal Service	Quality Control Analysis
		Near Vision	Static Strength	Wrist-Finger Speed
		Control Precision	Installation	Rate Control
39-0000	Personal Care And Service Occupations	Customer and Personal Service	Customer and Personal Service	Rate Control
		Oral Expression	Static Strength	Mathematics – Knowledge
		Oral Comprehension	Technology Design	Operation and Control
51-0000	Production Occupations	Production and Processing	Customer and Personal Service	Rate Control
		Near Vision	Technology Design	Operation and Control
		Problem Sensitivity	Installation	Quality Control Analysis
19-0000	Life, Physical, And Social Science Occupations	Written Comprehension	Science	Medicine and Dentistry
		Oral Comprehension	Technology Design	Rate Control
		Reading Comprehension	Operations Analysis	Operation and Control
11-0000	Management Occupations	Administration and Management	Science	Economics and Accounting
		Oral Expression	Philosophy and Theology	Medicine and Dentistry
		Oral Comprehension	Sociology and Anthropology	Mathematics – Knowledge
41-0000	Sales And Related Occupations	Customer and Personal Service	Customer and Personal Service	Economics and Accounting
		Oral Expression	Science	Mathematics – Knowledge
		Oral Comprehension	Technology Design	Rate Control
49-0000	Installation, Maintenance, And Repair Occupations	Mechanical	Installation	Operation and Control
		Near Vision	Customer and Personal Service	Rate Control
		Repairing	Technology Design	Quality Control Analysis
47-0000	Construction And Extraction Occupations	Building and Construction	Customer and Personal Service	Quality Control Analysis
		Mechanical	Foreign Language	Operation and Control
		Near Vision	Installation	Wrist-Finger Speed
25-0000	Education, Training, And Library Occupations	Education and Training	Science	Mathematics – Knowledge
		Oral Expression	Operations Analysis	Medicine and Dentistry
		English Language	Technology Design	Economics and Accounting

Table 6.9: For US sub-major occupation groups, ranked lists (the highest ranked, top, and lowest ranked, bottom) of O*NET features that are: currently high-valued; complementary and; anti-complementary.

SOC	Title	Current features	Complementary features	Anti-complementary features
43-0000	Office And Administrative Support Occupations	Customer and Personal Service	Service Orientation	Mathematics – Knowledge
		Oral Comprehension	Customer and Personal Service	Economics and Accounting
		Oral Expression	Technology Design	Rate Control
33-0000	Protective Service Occupations	Public Safety and Security	Customer and Personal Service	Rate Control
		Problem Sensitivity	Technology Design	Operation and Control
		English Language	Science	Quality Control Analysis
15-0000	Computer And Mathematical Occupations	Computers and Electronics	Science	Economics and Accounting
		Critical Thinking	Technology Design	Rate Control
		Problem Sensitivity	Design	Medicine and Dentistry
45-0000	Farming, Fishing, And Forestry Occupations	Static Strength	Customer and Personal Service	Rate Control
		Arm-Hand Steadiness	Static Strength	Wrist-Finger Speed
		Multilimb Coordination	Service Orientation	Operation and Control
23-0000	Legal Occupations	Oral Expression	Science	Economics and Accounting
		Law and Government	Sociology and Anthropology	Medicine and Dentistry
		English Language	Philosophy and Theology	Mathematics – Knowledge
37-0000	Building And Grounds Cleaning And Maintenance Occupations	Customer and Personal Service	Customer and Personal Service	Rate Control
		Trunk Strength	Static Strength	Wrist-Finger Speed
		English Language	Service Orientation	Operation and Control
17-0000	Architecture And Engineering Occupations	Engineering and Technology	Science	Operation and Control
		Mathematics – Knowledge	Technology Design	Rate Control
		Design	Operations Analysis	Medicine and Dentistry
31-0000	Healthcare Support Occupations	Customer and Personal Service	Customer and Personal Service	Rate Control
		Oral Comprehension	Technology Design	Mathematics – Knowledge
		English Language	Science	Computers and Electronics
35-0000	Food Preparation And Serving Related Occupations	Customer and Personal Service	Customer and Personal Service	Rate Control
		Oral Comprehension	Static Strength	Computers and Electronics
		Oral Expression	Service Orientation	Operation and Control
21-0000	Community And Social Service Occupations	Psychology	Operations Analysis	Medicine and Dentistry
		Therapy and Counseling	Science	Reaction Time
		Active Listening	Philosophy and Theology	Therapy and Counseling

6.5.6 New Occupations

It is also useful to think about the occupations which may emerge in the future in response to the drivers of labour market change we consider in our study. These occupations correspond to high-demand locations in the feature space and are not associated with existing occupations. The model allows us to identify a hypothetical occupation which is 'almost certain' (see Section 6.4.4 for a formal interpretation) to experience an increase in workforce share and the combination of skills, abilities and knowledge features most associated with it.

US

For the US, the model identifies four hypothetical occupations which would almost certainly experience a rise in demand. Table 6.10 ranks the top five O*NET features in declining order of feature value for each hypothetical occupation. (S) denotes that the variable is an O*NET skills feature, (K) is an O*NET knowledge feature and (A) is an O*NET abilities feature.

We can understand something about these hypothetical occupations by looking at existing occupations that are 'closest to them' (in declining order of proximity), as described in Figure 6.10. Of the twenty occupations presented here, eleven are defined by O*NET as enjoying a Bright Outlook and/or are expected to benefit from the growth of the green economy[9].

The employment time-series for these closest occupations is also plotted in Figure 6.11 for historical context.

[9]Specifically, Bright Outlook occupations are ones which are projected to grow much faster than average (employment increase of 14% or more) over the period 2014-2024; have 100,000 or more job openings over the period 2014-2024; or are new and emerging occupations in a high growth industry.

Table 6.10: The four new occupations found by our model for the US, as described by their top five O*NET features.

	Feature Rank				
Index	1st	2nd	3rd	4th	5th
1	Customer and Personal Service (K)	Static Strength (A)	Service Orientation (S)	Biology (K)	Arm-Hand Steadiness (A)
2	Building and Construction (K)	Customer and Personal Service (K)	Static Strength (A)	Manual Dexterity (A)	Arm-Hand Steadiness (A)
3	Engineering and Technology (K)	Science (S)	Written Comprehension (S)	Critical Thinking (S)	Design (K)
4	Education and Training (K)	Oral Comprehension (S)	Social Perceptiveness (S)	Written Comprehension (S)	Reading Comprehension (S)

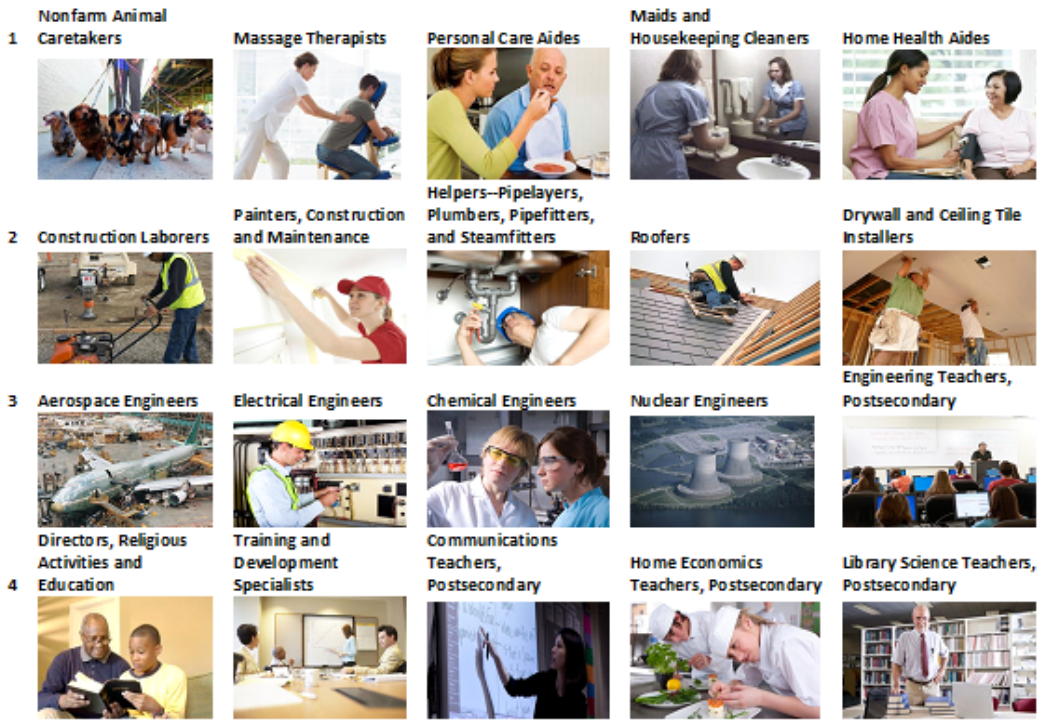
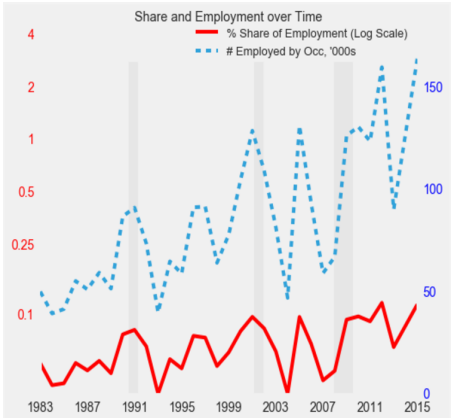
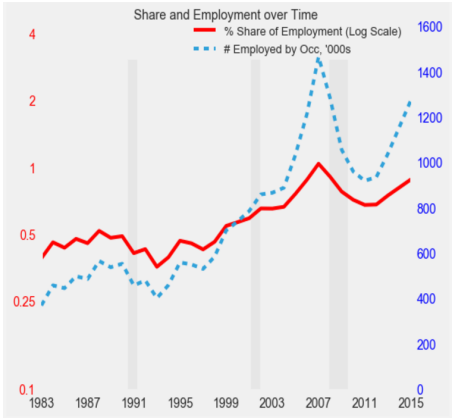


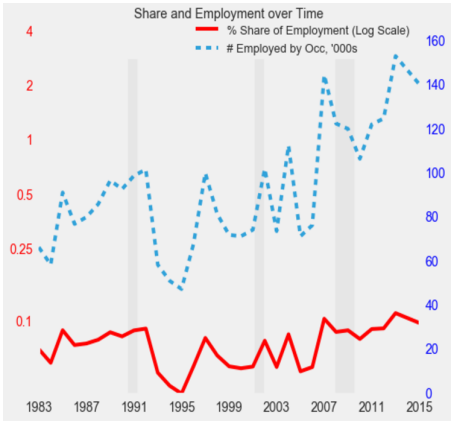
Figure 6.10: 'Closest' occupations to hypothetical new high demand occupations for the US.



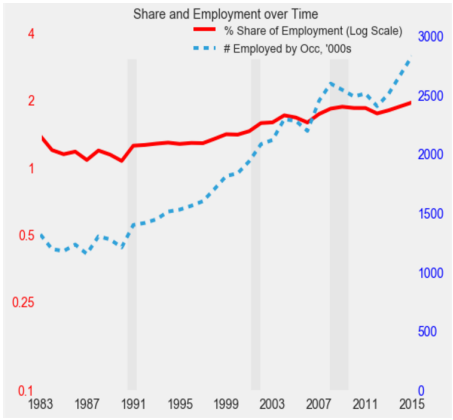
(a) Nonfarm Animal Caretakers



(b) Construction Laborers



(c) Aerospace Engineers



(d) Directors Religious Activities and Education

Figure 6.11: Time-series of employment for the 'closest' occupations to new US occupations, as tabulated in Figure 6.10.

These results provide another rejoinder to the view that jobs in the middle of the education and earnings distributions will disappear in the future. Two of the four occupations can be plausibly viewed as middle-skill jobs. Our first hypothetical occupation – which has similarities to social care work – is particularly interesting. On the one hand, it is a textbook example of a sector where the availability of low-skilled employees, the budgetary squeeze on government programs – Medicare and Medicaid account for roughly 70 percent of all long-term care dollars – and the legacy of the politics of race and gender have combined to create low-paid jobs with low status and precarious employment conditions (Duffy et al. 2015; Institute of Medicine (US) 2008). However the model points to bright demand prospects for care work which requires a mixture of tasks from across the skill

spectrum, including formal knowledge and training which, in principle, would support wage growth and job quality. Finally it is worth noting the extent to which interpersonal competencies feature across these hypothetical occupations.

6.6 Limitations and future work

While we believe that our research design has many appealing features which increases the usefulness of the findings compared with previous studies, we acknowledge there are important limitations. First, directional predictions may frustrate policymakers who seek more detailed information on which to base their decisions. Experimenting with a larger number of labels to achieve a finer distinction between different rates of change might have value in this respect, though we need to be mindful of the aforementioned cognitive limits associated with prediction over a 15-year horizon.

A second limitation is that we assess the implications for employment of only structural shifts in employer demand. In practice, however, employment opportunities will arise when workers retire from the workforce (or leave for other reasons) and need to be replaced. Indeed, replacement needs are expected to provide significantly more job openings than employment growth over the next decade (UK Commission for Employment and Skills 2014; U.S. Bureau of Labor Statistics 2016). Even those occupations where employer demand is otherwise expected to fall may still offer attractive career prospects. As such, incorporating estimates of the age structure of the workforce to predict replacement needs would complement our approach and assessment of future employment opportunities.

Third, it would be useful to understand more about the characteristics of jobs which are anticipated to become more important in coming years. Recent concerns that falling unemployment and the development of new business models have not been accompanied by the creation of 'good' jobs give this issue particular traction and timeliness (Taylor 2017). Earnings levels, career progression, working environment, job security, voice in organisational decisions among other things, provide objective and measurable benchmarks against which to assess job quality

(OECD 2016b). And in addition to the value that jobs have for the people who hold them, they also have potential spillovers, both positive and negative, on the rest of society which a full assessment would take into account.

A fourth avenue for developing our analysis would be to integrate trends more explicitly into the labelling process – for instance, to choose occupations which are most representative of the trends and likely to encourage reflection about them (as opposed to, or possibly combined with, using the active learning algorithm). Alternatively, workshop participants could be asked to rank the trends by their importance or relevance when labelling occupations as an input for our model, which would help sharpen interpretation of the results.

Finally, it would be useful to explore how estimates vary across countries (Hausmann et al. 2014; Beramendi et al. 2015). In the presence of cross-country variations in resources, institutions and technologies, even identical structural trends are likely to be channelled in different ways, which in turn would give rise to different labour market disruptions and opportunities.

6.7 Conclusions

Our study makes a significant contribution to the field of research focused on the future of skills and employment. Firstly, this is the first time that a principled Bayesian GP-HOR model has been used for this type of study. This model gives us the unique advantage of being able to consider participant confidence in data from our foresight workshops. It also allows us to make demand and skill predictions for unseen occupations whilst incorporating uncertainty.

Secondly, the foresight workshops employ a novel data collection methodology using active machine learning algorithms, which intelligently queries participants to maximise the informativeness of the data collected. Thirdly, the study employs an innovative approach to generating predictions about the future of skills, combining expert human judgement with machine learning techniques which can flexibly respond to natural patterns in the data. Our approach thus permits richer, more complex, non-linear interactions between variables – one which we exploit to

assess complementarities between skills and the implications for new occupations. Fourthly, the research is grounded in an explicit consideration of the diverse sources of structural change, any one of which can be expected to have major impacts on future employer skills needs. By making use of the detailed characterisation of occupations provided by the O*NET database, we are able to provide a higher resolution treatment of skills, knowledge types and abilities than is usually found in the skills literature. Finally, our research serves as a potentially important counterweight to the dominance of future automation in previous research.

In this chapter we show an application of the GP-HOR model developed in Chapter 3 using human-centric data in the form of workshop participant responses to questions about the future demand of occupations. Our skills results help inform policy makers about the future importance of 21st century skills - the combination of interpersonal and cognitive skills that has been an increasing preoccupation of policymakers in recent years.

7 Conclusion

Contents

7.1 Contributions

In this thesis we have studied how Gaussian Processes (GPs) can be used to improve bi-directional understanding between humans and machines, specifically: pairwise preference and heteroscedastic extensions to ordinal models in GPs, functional regression with GPs, and model interpretability with GPs. Applying the models developed in this thesis we investigated the Future of Skills in the economy. Here we briefly summarise the main contributions.

Chapter 3: Ordinal models In this chapter we addressed how ordinal models could be extended to include pairwise preferences and heteroscedasticity using GPs. We extended the capabilities of ordinal models in two ways using GPs, one being by introducing ordinal pairwise preferences and the other by introducing heteroscedastic ordinal labels. We developed the novel pairwise preference Gaussian Process Ordinal

Preference Learning (GP-OPL) model which extended the Gaussian Process Preference Learning (GP-PL) model (Chu and Ghahramani 2005d) to include preference strength. This is exemplified by statistically significant improvement of the median Kendall's tau metric relative to GP-PL, Rank Neural Network (RankNet) and Rank Support Vector Machine (RankSVM), on a range of synthetically derived datasets.

For the second extension of ordinal models we developed the novel Gaussian Process Heteroscedastic Ordinal Regression (GP-HOR) model, which extended Gaussian Process Ordinal Regression (GP-OR) (Chu and Ghahramani 2005a) to include ordinal heteroscedastic labels. Comparing our model to GP-OR experiments consistently showed that including confidence levels using GP-HOR increased the posterior density accuracy measured with the Wasserstein metric.

These two models give decision makers a greater understanding of human-centric data in the form of peoples' preferences due to the inclusion of self-reported confidence and ordinal graduation in relational preferences. Uncertainty of the model and data is captured naturally within our models due to our Bayesian framework. We have shown through experiments that our ordinal preference model is state-of-the-art in performance compared with other published models.

Chapter 4: Functional regression In this chapter we introduced the Gaussian Process Functional Generalized Additive Model (GP-FGAM) model, a novel GP model for non-linear functional regression, based on the Functional Generalized Additive Model (FGAM) model (Mathew W McLean et al. 2012). This probabilistic, non-parametric approach allows us to flexibly model fully observed functional inputs. The GP-FGAM model surpassed synthetic baselines in all cases measured with Root Mean Square Error (RMSE) and was able to correctly infer the latent surface as the Signal to Noise (SNR) value increased. The latent surface also gives decision makers valuable interpretable information about the salient areas within the functional predictor. The GP-FGAM model furthermore outperforms the FGAM model on a range of real-world data sets, whilst having the additional benefit of well

calibrated uncertainty estimates. Probability plots indicate that the model provides calibrated uncertainty estimates.

Chapter 5: Model interpretability This chapter addressed how to take advantage of GPs for model-agnostic interpretability of black-box models. A novel treatment of gradient averaging is presented using Bayesian quadrature resulting in the Principled Interpretability for Gradient Evaluation using Bayesian Quadrature (PIGEBaQ) model. Our method provides a fully principled way to average and interpret average function derivatives. In addition we provide an estimate for the marginal derivative, namely, the volatility of the underlying function. Our results show this to be a very useful tool for assessing movement in the sample.

Our synthetic experiments demonstrate that the quadrature model is able to, with high fidelity, reconstruct the true ranking of features based on their derivatives. The method was shown to be robust over a range of input dimensions and observed data points.

Applying our PIGEBaQ model to a pseudo-black-box model trained on real-world economic data, we extracted a ranked list of interpretable features, therefore, our principled methodology yields results that would help policy makers. This model gives decision makers the ability to probe a black-box model thus providing an ability to understand salient features of the problem, consequently helping to inform actions.

Chapter 6: Future of skills This chapter makes a significant contribution to the field of research focused on the future of skills and employment. We employed our GP-HOR model in order to quantitatively study the relationship between the constituents of jobs, namely, skills and demand for those skills. This model gives us the unique advantage of being able to consider participant confidence in data from our foresight workshops. It also allowed us to make demand and skill predictions for unseen occupations whilst incorporating uncertainty. Secondly, the foresight workshops employed a novel data collection methodology using active machine learning algorithms, which intelligently queried participants to maximise the informativeness of the data collected. Thirdly, the research employs an

innovative approach to generating predictions about the future of skills, combining expert human judgement with machine learning techniques which can flexibly respond to natural patterns in the data. Our approach thus permits richer, more complex, non-linear interactions between variables – one which we exploit to assess complementarities between skills and the implications for new occupations. Fourthly, the research is grounded in an explicit consideration of the diverse sources of structural change, any one of which can be expected to have major impacts on future employer skills needs. By making use of the detailed characterisation of occupations provided by the Occupational Information Network (O*NET) database, we are able to provide a higher resolution treatment of skills, knowledge types and abilities than is usually found in the skills literature. Finally, our research serves as a potentially important counterweight to the dominance of future automation in previous research.

7.2 Concluding thoughts

In all of the methods and algorithms presented in thesis there is scope for difference applications, as well as room for improvement.

The two ordinal methods presented in Chapter 3 have many potential applications in a range of fields such as psychology, multi-criteria decision analysis (used in business, government, and medicine (*Multiple-criteria decision analysis* 2020)), and others that require a confidence or strength of a pairwise preference or ordinal preference. The ordinal methods could also be extended to capture notions of non-transitivity, where a chain of pairwise preferences can loop back on itself.

The GP-FGAM model from Chapter 4 could enable industrialists and scientists to reason with the aid of uncertainty when regressing over functional data, such as other human-centric data e.g. Magnetic Resonance Imaging (MRI) or Magnetoencephalography (MEG) data. One could also imagine using the most important regions of the inferred latent surface of input function to inform future decision, e.g. In the context of the Tecator (TEC) dataset, costs could be reduced if only a narrow region of electromagnetic spectrum was needed for accurate regression.

Our PIGEBaQ model from Chapter 5 has considerable potential to illuminate what goes on in a wide range of black-box highly non-linear models. PIGEBaQ produces insights for local linear trends in data but also how volatile those trends are. This is critical in establishing confidence in machine learning methods where explanations matter.

The future of skills research in Chapter 6 was a novel approach to understanding the relationship between specific skills and demand for those skills but there are a range of improvements that could be made on a number of levels. The external O*NET dataset was created by asking workers how important a skill/knowledge/ability was for their occupation. Due to the resources needed to collect and curate responses the O*NET dataset is updated in segments leading to some sets of occupations having more up-to-date skill/knowledge/ability features compared to others. It would be helpful if O*NET was updated more often and more uniformly but, if this is not possible, it could be helpful to incorporate this temporal shift of skills into the model. On reflection our GP-HOR model could also be trained on each individual workshop participant with the results ensembled at the end rather than training one model for all individuals, this would have given more bi-modal distribution of demand. Of course in order to reduce some of the uncertainty in the model it would also be very helpful to increase the number of participants involved and occupations analysed.

Coming full circle back to the opening paragraphs of the Introduction chapter it is important to note that when two parties communicate effectively they can cooperate to do great things. Therefore, ensuring increased bi-directional understanding between humans and machines ultimately provides improved results. This thesis has made significant contributions to a number of methods which all enable better understanding between humans and machines. The interface between the two is occupied by human-centric data. As mentioned before human-centric data is full of nuance, and complexity, as well as being inherently uncertain. Throughout this thesis Gaussian Process (GP) models are used for their principled treatment of uncertainty. Probabilistic approaches will continue to be critical in capturing the

true preferences and intentions of people. Equally, as machine learning methods become more abstract and complex it is imperative that humans are kept in the loop, aware of how decisions that affect us are made.

Taking a step back we will also need to be mindful of the effect of the rapid automation of jobs has on not only the economy but society as a whole. Identifying vulnerable occupations and enabling workers in those occupations to reskill to alternative occupations will be critical in the times ahead.

Appendices

A

Model interpretability

Contents

A.1 Posterior Integral Equations

A.1.1 Kernel

The exponentiated quadratic kernel is:

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{x}')^T \Lambda^{-1} (\boldsymbol{x} - \boldsymbol{x}')}{2}\right) \tag{A.1}$$

The derivatives are

$$\frac{\partial}{\partial \boldsymbol{x}}\kappa(\boldsymbol{x},\boldsymbol{x}') = \sigma^2\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{x}')\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}')^T\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{x}')}{2}\right) \tag{A.2}$$

$$\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}\kappa(\boldsymbol{x},\boldsymbol{x}') = \sigma^2\Lambda^{-1}(\mathbb{I}-(\boldsymbol{x}-\boldsymbol{x})'(\boldsymbol{x}-\boldsymbol{x}')^T\Lambda^{-1})\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^T\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{x}')\right) \tag{A.3}$$

A.1.2 Posterior Mean of the Integral

The expected mean of the integral is:

$$\mathbb{E}[\boldsymbol{I}|f(\boldsymbol{x}_s)] = \sum_{i=1}^{N}\boldsymbol{z}^T K^{-1} f(\boldsymbol{x}_i), \tag{A.4}$$

where:

$$\boldsymbol{z}_n = \sum_{m=1}^{M} w_m \int K_{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{x}_n)p(\boldsymbol{x}|\theta_m)\mathrm{d}\boldsymbol{x} \tag{A.5}$$

Let us consider a single component:

$$\boldsymbol{z}_{nm} = \int K_{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{x}_n)p(\boldsymbol{x}|\theta_m)\mathrm{d}\boldsymbol{x} \tag{A.6}$$

$$= -\sigma^2\Lambda^{-1}\int(\boldsymbol{x}-\boldsymbol{x}_n')\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}_n)^T\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{x}_n)}{2}\right)\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m,\Sigma_m)\mathrm{d}\boldsymbol{x} \tag{A.7}$$

$$= -\sigma^2\Lambda^{-1}(2\pi)^{D/2}|\Lambda|^{1/2}\int(\boldsymbol{x}-\boldsymbol{x}_n)\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_n,\Lambda)\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m,\Sigma_m)\mathrm{d}\boldsymbol{x} \tag{A.8}$$

Using standard properties of products and integrals of multivariate Gaussians we arrive at:

$$\boldsymbol{z}_{nm} = -\frac{\sigma^2|\Lambda|^{1/2}}{|\Lambda+\Sigma_m|^{\frac{1}{2}}}(\Lambda+\Sigma_m)^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_m)\exp\left(-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_m)^T(\Lambda+\Sigma_m)^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_m)\right) \tag{A.9}$$

and consequently:

$$\boldsymbol{z}_n = \sum_{m=1}^{M}\frac{w_m\sigma^2|\Lambda|^{1/2}}{|\Lambda+\Sigma_m|^{\frac{1}{2}}}(\Lambda+\Sigma_m)^{-1}(\boldsymbol{\mu}_m-\boldsymbol{x}_n-)\exp\left(-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_m)^T(\Lambda+\Sigma_m)^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_m)\right). \tag{A.10}$$

A.1.3 Posterior Variance of the Integral

The posterior variance of the integral is given by:

$$\mathrm{Var}[\boldsymbol{I}|f(\boldsymbol{x}_s)] = \int\int K_{\boldsymbol{x},\boldsymbol{x}'}(\boldsymbol{x},\boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' - \boldsymbol{z}^T K^{-1}\boldsymbol{z}, \tag{A.11}$$

with $K_{\boldsymbol{x},\boldsymbol{x}'}(\boldsymbol{x},\boldsymbol{x}')$ the second derivative of the kernel function. The second term can be computed using z, the first expression is given by:

$$\left[\int\int K_{\boldsymbol{x},\boldsymbol{x}'}(\boldsymbol{x},\boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\right]_m \tag{A.12}$$

$$= \sigma^2\Lambda^{-1}(2\pi)^{D/2}|\Lambda|^{1/2}\int\int(I-(\boldsymbol{x}-\boldsymbol{x}')(\boldsymbol{x}-\boldsymbol{x}')^T)\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}',\Lambda)\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m,\Sigma_m)\mathcal{N}(\boldsymbol{x}'|\boldsymbol{\mu}_m,\Sigma_m)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \tag{A.13}$$

Exploiting the properties of the Gaussian distribution again and after some calculation we arrive at:

$$\int\int K_{\boldsymbol{x},\boldsymbol{x}'}(\boldsymbol{x},\boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \tag{A.14}$$

$$= \sum_{m=1}^{M} w_m \frac{\sigma^2|\Lambda|^{\frac{1}{2}}}{|\Lambda+2\Sigma_m|^{\frac{1}{2}}}(\Lambda^{-1}+(\Lambda+\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}+(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1} \tag{A.15}$$

$$+\Lambda^{-1}\Sigma_m(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}+2\Lambda^{-1}\boldsymbol{\mu}_m\boldsymbol{\mu}_m^T\Lambda^{-1} \tag{A.16}$$

$$+(\Lambda+\Sigma_m)^{-1}\Sigma_m(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\Sigma_m(\Lambda+\Sigma)^{-1} \tag{A.17}$$

$$+(\Lambda+\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\Sigma_m(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\Sigma_m(\Lambda+\Sigma_m)^{-1} \tag{A.18}$$

$$-(\Lambda_m+\Sigma_m)^{-1}\boldsymbol{\mu}_m\boldsymbol{\mu}_m^T\Lambda^{-1} \tag{A.19}$$

$$-(\Lambda+\Sigma_m)^{-1}\Lambda\Lambda^{-1}\Sigma_m(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\Lambda \tag{A.20}$$

$$-(\Lambda+\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\Sigma_m(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1} \tag{A.21}$$

$$-(\Lambda+\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\boldsymbol{\mu}_m\boldsymbol{\mu}_m^T\Lambda^{-1} \tag{A.22}$$

$$-\Lambda^{-1}\boldsymbol{\mu}_m\boldsymbol{\mu}_m^T(\Lambda+\Sigma_m)^{-1} \tag{A.23}$$

$$-\Sigma_m(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\Lambda(\Lambda+\Sigma_m)^{-1}\Lambda^{-1}\Lambda \tag{A.24}$$

$$-\Lambda^{-1}\Sigma_m(\Lambda+2\Sigma_m)^{-1}\Sigma_m\Lambda^{-1}\Sigma_m(\Lambda+\Sigma_m)^{-1} \tag{A.25}$$

$$-\Lambda^{-1}\boldsymbol{\mu}_m\boldsymbol{\mu}_m^T\Lambda^{-1}\Sigma_m(\Lambda+\Sigma_m)^{-1}) \tag{A.26}$$

A.2 Volatility

Let $\boldsymbol{x}^*$ be a prospective point with feature vector of dimensions $1 \times d$ and $\boldsymbol{X}$ a matrix of observations with dimension $n \times d$, where n is the number of observed points and d is the dimensionality of the feature vector. $\boldsymbol{y_X}$ is the target vector and has dimensions $n \times 1$.

Imagine taking our attention weighted derivative GP and marginalising out $\boldsymbol{x}^*$, this would give us a non-Gaussian distribution:

$$p(\boldsymbol{y'}) = \int_{\mathcal{X}} d\boldsymbol{x}^*\, p(\frac{\partial y}{\partial \boldsymbol{x}^*} \mid \boldsymbol{X}, y, \boldsymbol{x}^*) p\left(\boldsymbol{x}^*\right) = \int_{\mathcal{X}} d\boldsymbol{x}^*\, p(\boldsymbol{y'} \mid \boldsymbol{X}, y, \boldsymbol{x}^*) p\left(\boldsymbol{x}^*\right) \quad \text{(A.27)}$$

Where:

$$p(\boldsymbol{y'} \mid \boldsymbol{X}, y, \boldsymbol{x}^*) = \mathcal{N}(\boldsymbol{y'} \mid \boldsymbol{\mu_x}, \boldsymbol{\Sigma_x}) \quad \text{(A.28)}$$

$$\boldsymbol{\mu_x} = \boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)} \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})} \boldsymbol{y_X} \quad \text{(A.29)}$$

$$\boldsymbol{\Sigma_x} = \boldsymbol{K}_{\boldsymbol{y'}(\boldsymbol{x}^*),\boldsymbol{y'}(\boldsymbol{x}^*)} - \boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)} \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})} \boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)} \quad \text{(A.30)}$$

Here $p\left(\boldsymbol{x}^*\right)$ is taken to be a Gaussian Mixture Model:

$$p\left(\boldsymbol{x}^*\right) = \sum_p \pi_p\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu_p}, \boldsymbol{\Sigma_p}) \quad \text{(A.31)}$$

where π_p is the mixture coefficients and $\sum_p \pi_p = 1$, $\boldsymbol{\mu_p}$ and $\boldsymbol{\Sigma_p}$ are mixture means and covariances.

Defining the kernel notation:

$$\boldsymbol{K}_{A(\mathcal{C}),B(\mathcal{D})} = cov(A_{\mathcal{C}}, B_{\mathcal{D}}), \quad \text{(A.32)}$$

this reads: the covariance between functions A and B with input points to A being $\mathcal{C}$ and input points to B being $\mathcal{D}$.

A.3 Approximation of marginalised distribution

Approximating this distribution with a Gaussian distribution $q(\boldsymbol{y'}) = \mathcal{N}(\boldsymbol{y'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ we can minimise the $\min KL = \min KL\left(p(\boldsymbol{y'}) \mid\mid q(\boldsymbol{y'})\right)$, where:

$$\boldsymbol{\mu} = \mathbb{E}_{p(\boldsymbol{y'})}[\boldsymbol{y'}] \tag{A.33}$$

$$\boldsymbol{\Sigma} = \mathbb{V}ar_{p(\boldsymbol{y'})}[\boldsymbol{y'}] = \mathbb{E}_{p(\boldsymbol{y'})}[\boldsymbol{y'}\boldsymbol{y'}^T] - \mathbb{E}_{p(\boldsymbol{y'})}[\boldsymbol{y'}]\mathbb{E}_{p(\boldsymbol{y'})}[\boldsymbol{y'}]^T \tag{A.34}$$

A.3.1 Mean term μ

Expanding and rearranging the mean $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \int_{\mathcal{Y}'} d\boldsymbol{y'}\ \boldsymbol{y'} \int_{\mathcal{X}} d\boldsymbol{x}^*\ p(\boldsymbol{y'} \mid \boldsymbol{X}, y, \boldsymbol{x}^*) p\left(\boldsymbol{x}^*\right) \tag{A.35}$$

$$= \int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \int_{\mathcal{Y}'} d\boldsymbol{y'}\ \boldsymbol{y'}\ p(\boldsymbol{y'} \mid \boldsymbol{X}, y, \boldsymbol{x}^*) \tag{A.36}$$

$$= \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X}, y, \boldsymbol{x}^*)}[\boldsymbol{y'}]\right] \tag{A.37}$$

Substituting equation A.29 into the above:

$$\boldsymbol{\mu} = \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\boldsymbol{K}^T_{y(\boldsymbol{X}), \boldsymbol{y'}(\boldsymbol{x}^*)} \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}), y(\boldsymbol{X})} \boldsymbol{y}_{\boldsymbol{X}}\right] \tag{A.38}$$

$$= \int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \boldsymbol{K}^T_{y(\boldsymbol{X}), \boldsymbol{y'}(\boldsymbol{x}^*)} \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}), y(\boldsymbol{X})} \boldsymbol{y}_{\boldsymbol{X}} \tag{A.39}$$

$$= \left\{\int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \boldsymbol{K}^T_{y(\boldsymbol{X}), \boldsymbol{y'}(\boldsymbol{x}^*)}\right\} \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}), y(\boldsymbol{X})} \boldsymbol{y}_{\boldsymbol{X}} \tag{A.40}$$

Taking the bracketed section:

$$\int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \boldsymbol{K}^T_{y(\boldsymbol{X}), \boldsymbol{y'}(\boldsymbol{x}^*)} = \int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \left[\begin{matrix} \frac{\partial k(\boldsymbol{x}_1, \boldsymbol{x}_a)}{\partial \boldsymbol{x}_a} \\ \vdots \\ \frac{\partial k(\boldsymbol{x}_n, \boldsymbol{x}_a)}{\partial \boldsymbol{x}_a} \end{matrix}\right]_{\boldsymbol{x}_a = \boldsymbol{x}^*} \tag{A.41}$$

$$= \left[\begin{matrix} \int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \left.\frac{\partial k(\boldsymbol{x}_1, \boldsymbol{x}_a)}{\partial \boldsymbol{x}_a}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} \\ \vdots \\ \int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \left.\frac{\partial k(\boldsymbol{x}_n, \boldsymbol{x}_a)}{\partial \boldsymbol{x}_a}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} \end{matrix}\right] \tag{A.42}$$

Taking one element of this equation and substituting the RBF kernel:

$$\int_{\mathcal{X}} d\boldsymbol{x}^* \; p\left(\boldsymbol{x}^*\right) \left.\frac{\partial k(\boldsymbol{x}_l, \boldsymbol{x}_a)}{\partial \boldsymbol{x}_a}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} \tag{A.43}$$

$$= \int_{\mathcal{X}} d\boldsymbol{x}^* \; \sum_p \pi_p \, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu_p}, \boldsymbol{\Sigma_p}) k(\boldsymbol{x}_l, \boldsymbol{x}^*) \; (\boldsymbol{x}_l - \boldsymbol{x}^*) \boldsymbol{\Lambda}^{-1} \tag{A.44}$$

$$= \sum_p \pi_p \alpha_{\boldsymbol{\Lambda}} \int_{\mathcal{X}} d\boldsymbol{x}^* \, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu_p}, \boldsymbol{\Sigma_p}) \, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{x}_l, \boldsymbol{\Lambda}) \; (\boldsymbol{x}_l - \boldsymbol{x}^*) \boldsymbol{\Lambda}^{-1} \tag{A.45}$$

Combining the two multivariate Gaussian distributions:

$$\mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu_p}, \boldsymbol{\Sigma_p}) \, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{x}_l, \boldsymbol{\Lambda}) = k(l, p) \, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m1}) \tag{A.46}$$

where:

$$k(l, p) = \mathcal{N}(\boldsymbol{\mu_p} \mid \boldsymbol{x}_l, \boldsymbol{\Sigma_p} + \boldsymbol{\Lambda}) \tag{A.47}$$

$$\boldsymbol{\mu}_{m1} = (\boldsymbol{I} + \boldsymbol{\Sigma}_p \boldsymbol{\Lambda}^{-1})^{-1} \boldsymbol{\Sigma}_p (\boldsymbol{\Sigma_p}^{-1} \boldsymbol{\mu_p} + \boldsymbol{\Lambda}^{-1} \boldsymbol{x}_l) \tag{A.48}$$

$$= (\boldsymbol{I} + \boldsymbol{\Sigma}_p \boldsymbol{\Lambda}^{-1})^{-1} (\boldsymbol{\mu_p} + \boldsymbol{\Sigma}_p \boldsymbol{\Lambda}^{-1} \boldsymbol{x}_l) \tag{A.49}$$

$$\boldsymbol{\Sigma}_{m1} = (\boldsymbol{\Sigma_p}^{-1} + \boldsymbol{\Lambda}^{-1})^{-1} \tag{A.50}$$

$$= (\boldsymbol{I} + \boldsymbol{\Sigma}_p \boldsymbol{\Lambda}^{-1})^{-1} \boldsymbol{\Sigma}_p \tag{A.51}$$

Substituting back in and calculating:

$$\int_{\mathcal{X}} d\boldsymbol{x}^* \; p\left(\boldsymbol{x}^*\right) \left.\frac{\partial k(\boldsymbol{x}_l, \boldsymbol{x}_a)}{\partial \boldsymbol{x}_a}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} \tag{A.52}$$

$$= \sum_p \pi_p \alpha_{\boldsymbol{\Lambda}} k(l, p) \int_{\mathcal{X}} d\boldsymbol{x}^* \, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m1}) \; (\boldsymbol{x}_l - \boldsymbol{x}^*) \boldsymbol{\Lambda}^{-1} \tag{A.53}$$

$$= \sum_p \pi_p \alpha_{\boldsymbol{\Lambda}} k(l, p) \; (\boldsymbol{x}_l - \boldsymbol{\mu}_{m1}) \boldsymbol{\Lambda}^{-1} \tag{A.54}$$

A.3.2 Variance term Σ

Expanding the variance $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \mathbb{E}_{p(\boldsymbol{y'})}[\boldsymbol{y'}\boldsymbol{y'}^T] - \mathbb{E}_{p(\boldsymbol{y'})}[\boldsymbol{y'}]\mathbb{E}_{p(\boldsymbol{y'})}[\boldsymbol{y'}]^T \tag{A.55}$$

$$= \int_{\mathcal{Y}'} d\boldsymbol{y'}\ \boldsymbol{y'}\boldsymbol{y'}^T \int_{\mathcal{X}} d\boldsymbol{x}^*\ p(\boldsymbol{y'} \mid \boldsymbol{X}, y, \boldsymbol{x}^*)p\left(\boldsymbol{x}^*\right) - \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{A.56}$$

$$= \int_{\mathcal{X}} d\boldsymbol{x}^*\ p\left(\boldsymbol{x}^*\right) \int_{\mathcal{Y}'} d\boldsymbol{y'}\ \boldsymbol{y'}\boldsymbol{y'}^T\ p(\boldsymbol{y'} \mid \boldsymbol{X}, y, \boldsymbol{x}^*) - \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{A.57}$$

$$= \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}\boldsymbol{y'}^T]\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{A.58}$$

$$= \mathbb{E}_{p(\boldsymbol{x}^*)}[\mathbb{V}ar_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*))}[\boldsymbol{y'}] \tag{A.59}$$

$$+ \mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]^T] \tag{A.60}$$

$$- \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]\right]\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]\right]^T \tag{A.61}$$

Expanding the first term of $\boldsymbol{\Sigma}$:

$$\mathbb{V}ar_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}] + \mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]^T \tag{A.62}$$

$$= \boldsymbol{K}_{\boldsymbol{y'}(\boldsymbol{x}^*),\boldsymbol{y'}(\boldsymbol{x}^*)} - \boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)}\boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)} \tag{A.63}$$

$$+ \boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)}\boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})}\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)} \tag{A.64}$$

$$= \boldsymbol{K}_{\boldsymbol{y'}(\boldsymbol{x}^*),\boldsymbol{y'}(\boldsymbol{x}^*)} \tag{A.65}$$

$$+ \boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)}\left\{\boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})}\boldsymbol{y}_{\boldsymbol{X}}\boldsymbol{y}^T_{\boldsymbol{X}}\boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})} - \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})}\right\}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)} \tag{A.66}$$

$$= \boldsymbol{K}_{\boldsymbol{y'}(\boldsymbol{x}^*),\boldsymbol{y'}(\boldsymbol{x}^*)} + \boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)}\boldsymbol{A}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)} \tag{A.67}$$

where:

$$\boldsymbol{A} = \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})}\boldsymbol{y}_{\boldsymbol{X}}\boldsymbol{y}^T_{\boldsymbol{X}}\boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})} - \boldsymbol{K}^{-1}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})} \tag{A.68}$$

$$\boldsymbol{\Sigma} = \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\boldsymbol{K}_{\boldsymbol{y'}(\boldsymbol{x}^*),\boldsymbol{y'}(\boldsymbol{x}^*)} + \boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)}\boldsymbol{A}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y'}(\boldsymbol{x}^*)}\right] \tag{A.69}$$

$$- \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]\right]\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\mathbb{E}_{p(\boldsymbol{y'}|\boldsymbol{X},y,\boldsymbol{x}^*)}[\boldsymbol{y'}]\right]^T \tag{A.70}$$

The first term in equation A.69 can be calculated straight away:

$$\mathbb{E}_{p(\boldsymbol{x}^*)}[\boldsymbol{K}_{\boldsymbol{y}'(\boldsymbol{x}^*),\boldsymbol{y}'(\boldsymbol{x}^*)}] = \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\left.\frac{\partial^2 k(\boldsymbol{x}_a,\boldsymbol{x}_b)}{\partial \boldsymbol{x}_a \partial \boldsymbol{x}_b}\right|_{\boldsymbol{x}_a=\boldsymbol{x}_b=\boldsymbol{x}^*}\right] \tag{A.71}$$

$$= \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_f \boldsymbol{\Lambda}^{-1}\right] \tag{A.72}$$

$$= \sigma_f \boldsymbol{\Lambda}^{-1} \tag{A.73}$$

The second term in equation A.69 needs to be teased apart:

$$\left\{\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y}'(\boldsymbol{x}^*)}\boldsymbol{A}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'(\boldsymbol{x}^*)}\right]\right\}_{i,j} = \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y}'_i(\boldsymbol{x}^*)}\boldsymbol{A}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'_j(\boldsymbol{x}^*)}\right] \tag{A.74}$$

where $\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'_i(\boldsymbol{x}^*)}$ takes the derivative only on the ith feature and is therefore a column vector. Using the useful result:

$$\mathbb{E}[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}] = \mathrm{Tr}[\boldsymbol{A}\mathbb{E}[\boldsymbol{y}\boldsymbol{x}^T]] \tag{A.75}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are stochastic vectors with dimensions $n \times 1$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the element-by-element result of equation A.74 can be calculated.

Equation A.74 becomes:

$$\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y}'_i(\boldsymbol{x}^*)}\boldsymbol{A}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'_j(\boldsymbol{x}^*)}\right] \tag{A.76}$$

$$= \mathrm{Tr}\left[\boldsymbol{A}\mathbb{E}_{p(\boldsymbol{x}^*)}[\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'_j(\boldsymbol{x}^*)}\boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y}'_i(\boldsymbol{x}^*)}]\right] \tag{A.77}$$

Taking out the expectation term:

$$\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'_j(\boldsymbol{x}^*)}\boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y}'_i(\boldsymbol{x}^*)}\right] \tag{A.78}$$

$$= \int d\boldsymbol{x}^* p(\boldsymbol{x}^*) \left[\begin{matrix}\frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,j}} \\ \vdots \\ \frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,j}}\end{matrix}\right]_{\boldsymbol{x}_a=\boldsymbol{x}^*} \left[\begin{matrix}\frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,i}} \\ \vdots \\ \frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,i}}\end{matrix}\right]^T_{\boldsymbol{x}_a=\boldsymbol{x}^*} \tag{A.79}$$

$$= \int d\boldsymbol{x}^* p(\boldsymbol{x}^*) \left[\begin{matrix}\frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,j}}\frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,i}} & \cdots & \frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,j}}\frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,i}} \\ \vdots & \ddots & \vdots \\ \frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,j}}\frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,i}} & \cdots & \frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,j}}\frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,i}}\end{matrix}\right]_{\boldsymbol{x}_a=\boldsymbol{x}^*} \tag{A.80}$$

Looking at one element of this $n \times n$ matrix:

$$\int d\boldsymbol{x}^* \left.\frac{\partial k(\boldsymbol{x}_l, \boldsymbol{x}_a)}{\partial x_{a,j}}\right|_{\boldsymbol{x}_a=\boldsymbol{x}^*} \left.\frac{\partial k(\boldsymbol{x}_m, \boldsymbol{x}_a)}{\partial x_{a,i}}\right|_{\boldsymbol{x}_a=\boldsymbol{x}^*} p\left(\boldsymbol{x}^*\right) \tag{A.81}$$

$$= \int d\boldsymbol{x}^* k(\boldsymbol{x}_l, \boldsymbol{x}^*)\; k(\boldsymbol{x}_m, \boldsymbol{x}^*)(\boldsymbol{x}_l - \boldsymbol{x}^*)\boldsymbol{\Lambda}_j^{-1}(\boldsymbol{x}_m - \boldsymbol{x}^*)\boldsymbol{\Lambda}_i^{-1} p\left(\boldsymbol{x}^*\right) \tag{A.82}$$

$$= \int d\boldsymbol{x}^* k(\boldsymbol{x}_l, \boldsymbol{x}^*)\; k(\boldsymbol{x}_m, \boldsymbol{x}^*)(\boldsymbol{\Lambda}_j^{-1})^T(\boldsymbol{x}_l - \boldsymbol{x}^*)^T(\boldsymbol{x}_m - \boldsymbol{x}^*)\boldsymbol{\Lambda}_i^{-1} p\left(\boldsymbol{x}^*\right) \tag{A.83}$$

Since $k(\boldsymbol{x}_l, \boldsymbol{x}^*)$, $k(\boldsymbol{x}_m, \boldsymbol{x}^*)$ and $p\left(\boldsymbol{x}^*\right)$ are all (scaled) normal distributions dependent on $\boldsymbol{x}^*$, they can be combined into one normal distribution dependent on $\boldsymbol{x}^*$ and two other normal distribution not dependent on $\boldsymbol{x}^*$.

Substituting the forms for all three distributions:

$$k(\boldsymbol{x}_l, \boldsymbol{x}^*)\; k(\boldsymbol{x}_m, \boldsymbol{x}^*)\; p\left(\boldsymbol{x}^*\right) \tag{A.84}$$

$$= k(\boldsymbol{x}^*, \boldsymbol{x}_l)\; k(\boldsymbol{x}^*, \boldsymbol{x}_m)\; p\left(\boldsymbol{x}^*\right) \tag{A.85}$$

$$= \alpha_{\boldsymbol{\Lambda}}\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{x}_l, \boldsymbol{\Lambda}) \alpha_{\boldsymbol{\Lambda}}\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{x}_m, \boldsymbol{\Lambda}) \sum_p \pi_p\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu}_{\boldsymbol{p}}, \boldsymbol{\Sigma}_{\boldsymbol{p}}) \tag{A.86}$$

$$= \sum_p \pi_p\, \alpha_{\boldsymbol{\Lambda}}^2\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{x}_l, \boldsymbol{\Lambda})\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{x}_m, \boldsymbol{\Lambda})\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu}_{\boldsymbol{p}}, \boldsymbol{\Sigma}_{\boldsymbol{p}}) \tag{A.87}$$

$$= \sum_p \pi_p\, \alpha_{\boldsymbol{\Lambda}}^2 c_1(l, m)\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu}_{\boldsymbol{c_1}}, \boldsymbol{\Sigma}_{\boldsymbol{c_1}})\, \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu}_{\boldsymbol{p}}, \boldsymbol{\Sigma}_{\boldsymbol{p}}) \tag{A.88}$$

$$= \sum_p \pi_p\, \alpha_{\boldsymbol{\Lambda}}^2 c_1(l, m) c_2(l, m, p) \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu}_{\boldsymbol{c_2}}, \boldsymbol{\Sigma}_{\boldsymbol{c_2}}) \tag{A.89}$$

where c_1 terms are:

$$c_1(l, m) = \mathcal{N}(\boldsymbol{x}_l \mid \boldsymbol{x}_m, 2\boldsymbol{\Lambda}) \tag{A.90}$$

$$\boldsymbol{\mu}_{\boldsymbol{c_1}} = \tfrac{1}{2}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}^{-1}\boldsymbol{x}_l + \boldsymbol{\Lambda}^{-1}\boldsymbol{x}_m) \tag{A.91}$$

$$= \tfrac{1}{2}(\boldsymbol{x}_l + \boldsymbol{x}_m) \tag{A.92}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{c_1}} = (\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Lambda}^{-1})^{-1} \tag{A.93}$$

$$= \tfrac{1}{2}\boldsymbol{\Lambda} \tag{A.94}$$

where c_2 terms are:

$$c_2(n,m,p) = \mathcal{N}\left(\tfrac{1}{2}(\boldsymbol{x}_n + \boldsymbol{x}_m) \mid \boldsymbol{\mu_p}, \tfrac{1}{2}\boldsymbol{\Lambda} + \boldsymbol{\Sigma_p}\right) \tag{A.95}$$

$$\boldsymbol{\mu_{c_2}} = \boldsymbol{\Lambda}(2\boldsymbol{\Sigma_p} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Sigma_p}\left((\tfrac{1}{2}\boldsymbol{\Lambda})^{-1}\tfrac{1}{2}(\boldsymbol{x}_n + \boldsymbol{x}_m) + \boldsymbol{\Sigma_p}^{-1}\boldsymbol{\mu_p}\right) \tag{A.96}$$

$$= \boldsymbol{\Lambda}(2\boldsymbol{\Sigma_p} + \boldsymbol{\Lambda})^{-1}\left(\boldsymbol{\Sigma_p}\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}_n + \boldsymbol{x}_m) + \boldsymbol{\mu_p}\right) \tag{A.97}$$

$$\boldsymbol{\Sigma_{c_2}} = ((\tfrac{1}{2}\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma_p}^{-1})^{-1} \tag{A.98}$$

$$= (2\boldsymbol{\Sigma_p}\boldsymbol{\Lambda}^{-1} + \boldsymbol{I})^{-1}\boldsymbol{\Sigma_p} \tag{A.99}$$

$$= \boldsymbol{\Lambda}(2\boldsymbol{\Sigma_p} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Sigma_p} \tag{A.100}$$

Substituting equation A.89 into A.83:

$$\int d\boldsymbol{x}^* \left.\frac{\partial k(\boldsymbol{x}_l, \boldsymbol{x}_a)}{\partial x_{a,j}}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} \left.\frac{\partial k(\boldsymbol{x}_m, \boldsymbol{x}_a)}{\partial x_{a,i}}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} p(\boldsymbol{x}^*) \tag{A.101}$$

$$= \int d\boldsymbol{x}^* \sum_p \pi_p \, \alpha_{\boldsymbol{\Lambda}}^2 c_1(l,m)c_2(l,m,p) \tag{A.102}$$

$$\mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu_{c_2}}, \boldsymbol{\Sigma_{c_2}})(\boldsymbol{\Lambda}_j^{-1})^T(\boldsymbol{x}_l - \boldsymbol{x}^*)^T(\boldsymbol{x}_m - \boldsymbol{x}^*)\boldsymbol{\Lambda}_i^{-1}$$

$$= \sum_p \pi_p \, \alpha_{\boldsymbol{\Lambda}}^2 c_1(l,m)c_2(l,m,p) \tag{A.103}$$

$$(\boldsymbol{\Lambda}_j^{-1})^T \left[\int d\boldsymbol{x}^* \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu_{c_2}}, \boldsymbol{\Sigma_{c_2}})(\boldsymbol{x}_l - \boldsymbol{x}^*)^T(\boldsymbol{x}_m - \boldsymbol{x}^*)\right] \boldsymbol{\Lambda}_i^{-1}$$

$$= \sum_p \pi_p \, \alpha_{\boldsymbol{\Lambda}}^2 c_1(l,m)c_2(l,m,p) \tag{A.104}$$

$$(\boldsymbol{\Lambda}_j^{-1})^T \left[\int d\boldsymbol{x}^* \mathcal{N}(\boldsymbol{x}^* \mid \boldsymbol{\mu_{c_2}}, \boldsymbol{\Sigma_{c_2}})(\boldsymbol{x}_l^T\boldsymbol{x}_m - \boldsymbol{x}_l^T\boldsymbol{x}^* - \boldsymbol{x}^{*T}\boldsymbol{x}_m + \boldsymbol{x}^{*T}\boldsymbol{x}^*)\right] \boldsymbol{\Lambda}_i^{-1}$$

Calculating the integral within the square brackets:

$$\int d\boldsymbol{x}^* \left.\frac{\partial k(\boldsymbol{x}_l, \boldsymbol{x}_a)}{\partial x_{a,j}}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} \left.\frac{\partial k(\boldsymbol{x}_m, \boldsymbol{x}_a)}{\partial x_{a,i}}\right|_{\boldsymbol{x}_a = \boldsymbol{x}^*} p(\boldsymbol{x}^*) \tag{A.105}$$

$$= \sum_p \pi_p \, \alpha_{\boldsymbol{\Lambda}}^2 c_1(l,m)c_2(l,m,p) \tag{A.106}$$

$$(\boldsymbol{\Lambda}_j^{-1})^T \left[\boldsymbol{x}_l^T\boldsymbol{x}_m - \boldsymbol{x}_l^T\boldsymbol{\mu_{c_2}} - \boldsymbol{\mu_{c_2}}^T\boldsymbol{x}_m + \boldsymbol{\Sigma_{c_2}} + \boldsymbol{\mu_{c_2}}^T\boldsymbol{\mu_{c_2}}\right] \boldsymbol{\Lambda}_i^{-1}$$

Since the terms $\boldsymbol{x}_l$, $\boldsymbol{x}_m$, $\boldsymbol{\mu_{c_2}}$, $\boldsymbol{\Sigma_{c_2}}$, $\pi_p \, \alpha_{\boldsymbol{\Lambda}}^2 c_1(l,m)c_2(l,m,p)$ are all independent of i,j they can be calculated independently to produce a 4-D tensor for all l,m with dimensions $n \times n \times d \times d$. The terms $(\boldsymbol{\Lambda}_i^{-1})^T$ and $\boldsymbol{\Lambda}_j^{-1}$ for all i,j can be

broadcast multiplied (along the i, j dimensions) with this 4-D tensor to produce another 4-D tensor with similar dimensions.

Recalling equation A.77:

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y}'_i(\boldsymbol{x}^*)}\boldsymbol{A}\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'_j(\boldsymbol{x}^*)}\right] \\ = \mathrm{Tr}\left[\boldsymbol{A}\mathbb{E}_{p(\boldsymbol{x}^*)}[\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'_j(\boldsymbol{x}^*)}\boldsymbol{K}^T_{y(\boldsymbol{X}),\boldsymbol{y}'_i(\boldsymbol{x}^*)}]\right]\end{aligned}$$

The 2-D matrix $\boldsymbol{A}$ can be broadcast multiplied with the 4-D tensor created using equation A.106 along the l, m dimensions. Taking the trace over the l, m dimensions of the resultant 4-D tensor reduces to a 2-D tensor with dimensions $d \times d$.

A.4 Kernel definitions

Define all Gram matrices from section A.2:

$$\boldsymbol{K}_{\boldsymbol{y}(\boldsymbol{X}),y(\boldsymbol{X})} = \begin{bmatrix} k(\boldsymbol{x}_1,\boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_1,\boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n,\boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_n,\boldsymbol{x}_n) \end{bmatrix} \tag{A.107}$$

$$\boldsymbol{K}_{y(\boldsymbol{X}),\boldsymbol{y}'(\boldsymbol{x}^*)} = \begin{bmatrix} \frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial \boldsymbol{x}_a} \\ \vdots \\ \frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial \boldsymbol{x}_a} \end{bmatrix}_{\boldsymbol{x}_a=\boldsymbol{x}^*} = \begin{bmatrix} \frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,1}} & \dots & \frac{\partial k(\boldsymbol{x}_1,\boldsymbol{x}_a)}{\partial x_{a,d}} \\ \vdots & \ddots & \vdots \\ \frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,1}} & \dots & \frac{\partial k(\boldsymbol{x}_n,\boldsymbol{x}_a)}{\partial x_{a,d}} \end{bmatrix}_{\boldsymbol{x}_a=\boldsymbol{x}^*} \tag{A.108}$$

$$\boldsymbol{K}_{\boldsymbol{y}'(\boldsymbol{x}^*),\boldsymbol{y}'(\boldsymbol{x}^*)} = \left.\frac{\partial^2 k(\boldsymbol{x}_a,\boldsymbol{x}_b)}{\partial \boldsymbol{x}_a \partial \boldsymbol{x}_b}\right|_{\boldsymbol{x}_a=\boldsymbol{x}_b=\boldsymbol{x}^*} = \begin{bmatrix} \frac{\partial^2 k(\boldsymbol{x}_a,\boldsymbol{x}_b)}{\partial x_{a,1}\partial x_{b,1}} & \dots & \frac{\partial^2 k(\boldsymbol{x}_a,\boldsymbol{x}_b)}{\partial x_{a,1}\partial x_{b,d}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 k(\boldsymbol{x}_a,\boldsymbol{x}_b)}{\partial x_{a,d}\partial x_{b,1}} & \dots & \frac{\partial^2 k(\boldsymbol{x}_a,\boldsymbol{x}_b)}{\partial x_{a,d}\partial x_{b,d}} \end{bmatrix}_{\boldsymbol{x}_a=\boldsymbol{x}_b=\boldsymbol{x}^*} \tag{A.109}$$

A.4.1 Square exponential kernel

Defining the kernel and derivatives:

Square exponential kernel:

$$k(\boldsymbol{x}_a,\boldsymbol{x}_b) = \sigma_f \exp\left(-\frac{1}{2}(\boldsymbol{x}_a-\boldsymbol{x}_b)\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}_a-\boldsymbol{x}_b)^T\right) \tag{A.110}$$

$$= \alpha_{\boldsymbol{\Lambda}}\,\mathcal{N}(\boldsymbol{x}_a \mid \boldsymbol{x}_b, \boldsymbol{\Lambda}) \tag{A.111}$$

where $\alpha_{\mathbf{\Lambda}} = \sigma_f |2\pi\mathbf{\Lambda}|^{\frac{1}{2}}$

First derivative:

$$\frac{\partial k(\boldsymbol{x}_a, \boldsymbol{x}_b)}{\partial \boldsymbol{x}_b} = k(\boldsymbol{x}_a, \boldsymbol{x}_b)\ (\boldsymbol{x}_a - \boldsymbol{x}_b)\mathbf{\Lambda}^{-1} \tag{A.112}$$

First derivative for one dimension:

$$\frac{\partial k(\boldsymbol{x}_a, \boldsymbol{x}_b)}{\partial x_{b,i}} = k(\boldsymbol{x}_a, \boldsymbol{x}_b)\ (\boldsymbol{x}_a - \boldsymbol{x}_b)\mathbf{\Lambda}_i^{-1} \tag{A.113}$$

where $\mathbf{\Lambda}_i^{-1}$ is the ith column of $\mathbf{\Lambda}^{-1}$.

Second derivative:

$$\frac{\partial^2 k(\boldsymbol{x}_a, \boldsymbol{x}_b)}{\partial \boldsymbol{x}_a \partial \boldsymbol{x}_b} = k(\boldsymbol{x}_a, \boldsymbol{x}_b)\ \left(\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}(\boldsymbol{x}_a - \boldsymbol{x}_b)^T(\boldsymbol{x}_a - \boldsymbol{x}_b)\mathbf{\Lambda}^{-1}\right) \tag{A.114}$$

B
Future of Skills

Contents

B.1 Skills

US

Table B.1: A ranking, by average derivative, of the importance of O*NET variables to future demand for US occupations.

Rank	O*NET Variable	Class	Average Derivative
1	Customer and Personal Service	Knowledge	2.578
2	Technology Design	Skills	2.565
3	Science	Skills	2.557
4	Service Orientation	Skills	2.229
5	Education and Training	Knowledge	2.087
6	Static Strength	Abilities	1.965
7	Philosophy and Theology	Knowledge	1.953
8	Instructing	Skills	1.847
9	Installation	Skills	1.843
10	Sociology and Anthropology	Knowledge	1.655
11	Fluency of Ideas	Abilities	1.602

Table B.1: A ranking, by average derivative, of the importance of O*NET variables to future demand for US occupations.

Rank	O*NET Variable	Class	Average Derivative
12	Stamina	Abilities	1.570
13	Personnel and Human Resources	Knowledge	1.544
14	Complex Problem Solving	Skills	1.377
15	Management of Material Resources	Skills	1.227
16	Extent Flexibility	Abilities	1.226
17	Operations Analysis	Skills	1.189
18	Design	Knowledge	1.170
19	Equipment Selection	Skills	1.162
20	Psychology	Knowledge	1.071
21	Dynamic Strength	Abilities	1.067
22	Originality	Abilities	1.048
23	Management of Personnel Resources	Skills	1.041
24	Chemistry	Knowledge	1.040
25	Therapy and Counseling	Knowledge	1.016
26	Foreign Language	Knowledge	1.012
27	Arm-Hand Steadiness	Abilities	1.008
28	Learning Strategies	Skills	0.985
29	Physics	Knowledge	0.971
30	Active Learning	Skills	0.940
31	Memorization	Abilities	0.914
32	Administration and Management	Knowledge	0.902
33	Dynamic Flexibility	Abilities	0.844
34	Time Sharing	Abilities	0.841
35	Social Perceptiveness	Skills	0.745
36	Writing	Skills	0.737
37	Manual Dexterity	Abilities	0.721
38	Sound Localization	Abilities	0.659
39	Multilimb Coordination	Abilities	0.652
40	Gross Body Coordination	Abilities	0.634
41	Engineering and Technology	Knowledge	0.631
42	Speaking	Skills	0.622
43	Reading Comprehension	Skills	0.580
44	Trunk Strength	Abilities	0.552
45	Geography	Knowledge	0.533
46	Communications and Media	Knowledge	0.527

Table B.1: A ranking, by average derivative, of the importance of O*NET variables to future demand for US occupations.

Rank	O*NET Variable	Class	Average Derivative
47	Telecommunications	Knowledge	0.514
48	Speech Recognition	Abilities	0.510
49	Information Ordering	Abilities	0.454
50	Inductive Reasoning	Abilities	0.441
51	Active Listening	Skills	0.391
52	Coordination	Skills	0.379
53	Depth Perception	Abilities	0.351
54	Far Vision	Abilities	0.348
55	Mechanical	Knowledge	0.341
56	Written Comprehension	Abilities	0.332
57	Problem Sensitivity	Abilities	0.330
58	Monitoring	Skills	0.267
59	Time Management	Skills	0.210
60	Deductive Reasoning	Abilities	0.171
61	Written Expression	Abilities	0.162
62	History and Archeology	Knowledge	0.160
63	Visual Color Discrimination	Abilities	0.155
64	Finger Dexterity	Abilities	0.142
65	Glare Sensitivity	Abilities	0.091
66	Judgment and Decision Making	Skills	0.069
67	Oral Expression	Abilities	0.050
68	Peripheral Vision	Abilities	0.046
69	Visualization	Abilities	0.043
70	Persuasion	Skills	0.034
71	Gross Body Equilibrium	Abilities	0.012
72	Oral Comprehension	Abilities	0.012
73	Spatial Orientation	Abilities	-0.053
74	Public Safety and Security	Knowledge	-0.081
75	Explosive Strength	Abilities	-0.103
76	Management of Financial Resources	Skills	-0.138
77	Critical Thinking	Skills	-0.176
78	Programming	Skills	-0.182
79	Speech Clarity	Abilities	-0.299
80	Speed of Limb Movement	Abilities	-0.326
81	Speed of Closure	Abilities	-0.328

Table B.1: A ranking, by average derivative, of the importance of O*NET variables to future demand for US occupations.

Rank	O*NET Variable	Class	Average Derivative
82	Transportation	Knowledge	-0.365
83	Troubleshooting	Skills	-0.367
84	Systems Analysis	Skills	-0.391
85	Selective Attention	Abilities	-0.424
86	Sales and Marketing	Knowledge	-0.434
87	Near Vision	Abilities	-0.440
88	Category Flexibility	Abilities	-0.517
89	Negotiation	Skills	-0.559
90	Equipment Maintenance	Skills	-0.561
91	Systems Evaluation	Skills	-0.572
92	Clerical	Knowledge	-0.601
93	Night Vision	Abilities	-0.701
94	Repairing	Skills	-0.715
95	Response Orientation	Abilities	-0.737
96	Auditory Attention	Abilities	-0.822
97	Operation Monitoring	Skills	-0.910
98	Flexibility of Closure	Abilities	-0.924
99	Hearing Sensitivity	Abilities	-0.944
100	Mathematics – Skills	Skills	-0.944
101	Law and Government	Knowledge	-0.949
102	Mathematical Reasoning	Abilities	-1.024
103	English Language	Knowledge	-1.079
104	Medicine and Dentistry	Knowledge	-1.233
105	Number Facility	Abilities	-1.399
106	Reaction Time	Abilities	-2.014
107	Quality Control Analysis	Skills	-2.027
108	Economics and Accounting	Knowledge	-2.043
109	Computers and Electronics	Knowledge	-2.052
110	Wrist-Finger Speed	Abilities	-2.053
111	Operation and Control	Skills	-2.334
112	Mathematics – Knowledge	Knowledge	-2.365
113	Rate Control	Abilities	-2.684

References

Abramowitz, Milton, Irene A. Stegun, and David Miller (1965). *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables (National Bureau of Standards Applied Mathematics Series No. 55)*. arXiv: 1701.01870.

Aguilera, Ana M et al. (2013). "Functional Analysis of Chemometric Data". In: *Open Journal of Statistics* 3, pp. 334–343.

Alcorn, Michael A. (2019). *Keras RankNet Implementation*. original-date: 2017-04-13T13:58:27Z. URL: https://github.com/airalcorn2/RankNet.

Arnborg, S. and G. Sjödin (2001). "On the foundations of Bayesianism". In: *AIP Conference Proceedings* 568.1, pp. 61–71. URL: https://aip.scitation.org/doi/abs/10.1063/1.1381871.

Arntz, M., T. Gregory, and U. Zierahn (2016). *The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis*. Tech. rep. 189. OECD Social, Employment and Migration Working Papers.

Atkinson, Robert D. and John Wu (2017). *False Alarmism: Technological Disruption and the U.S. Labor Market, 1850–2015*. Tech. rep. Information Technology and Innovation Foundation.

Babibker, Housam and Randy Goebel (2017). "Using KL-divergence to focus Deep Visual Explanation". In:

Baehrens, David et al. (2010). "How to Explain Individual Classification Decisions". In: *Journal of Machine Learning Research* 11, pp. 1803–1831. arXiv: 0912.1128. URL: https://is.tuebingen.mpg.de/fileadmin/user%5C_upload/files/publications/baehrens10a%5C_[0].pdf.

Bakhshi, Hasan, Jonathan M Downing, et al. (2017). *The future of skills: employment in 2030*. Pearson.

Bakhshi, Hasan, Jonathan M. Downing, et al. (2017). *The future of skills: employment in 2030*. OCLC: 1005079628. United Kingdom: Pearson.

Barratt, Shane (2017). "InterpNET: Neural Introspection for Interpretable Deep Learning". In: arXiv: 1710.09511.

Bellemare, Marc G. et al. (2017). "The Cramer Distance as a Solution to Biased Wasserstein Gradients". In: *arXiv:1705.10743 [cs, stat]*. arXiv: 1705.10743. URL: http://arxiv.org/abs/1705.10743.

Beramendi, Pablo et al. (2015). *The Politics of Advanced Capitalism*. en. Google-Books-ID: 9Q3UBwAAQBAJ. Cambridge University Press.

Bergstra, James S. et al. (2011). "Algorithms for Hyper-Parameter Optimization". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., pp. 2546–2554. URL: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf.

Boyd, John P. (1992). "Defeating the Runge phenomenon for equispaced polynomial interpolation via Tikhonov regularization". In: *Applied Mathematics Letters* 5.6,

pp. 57–59. URL: http://www.sciencedirect.com/science/article/pii/089396599290014Z.

Bromiley, P A. "Products and Convolutions of Gaussian Probability Density Functions". In: p. 13.

Burges, Chris et al. (2005). "Learning to Rank Using Gradient Descent". In: *Proceedings of the 22Nd International Conference on Machine Learning*. ICML '05. New York, NY, USA: ACM, pp. 89–96. URL: http://doi.acm.org/10.1145/1102351.1102363.

Byrd, Richard H et al. (1995). "A limited memory algorithm for bound constrained optimization". In: *SIAM Journal on Scientific Computing* 16.5, pp. 1190–1208.

Cardot, Hervé and Pacal Sarda (2005). "Estimation in generalized linear models for functional data via penalized likelihood". In: *Journal of Multivariate Analysis* 92.1, pp. 24–41.

Chen, Dong, Peter Hall, and Hans Georg Müller (2011). "Single and multiple index functional regression models with nonparametric link". In: *Annals of Statistics* 39.3, pp. 1720–1747. arXiv: arXiv:1211.5018v1.

Choi, Edward et al. (2016). "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism". In: *NIPS*. arXiv: 1608.05745. URL: http://papers.nips.cc/paper/6321-retain-an-interpretable-predictive-model-for-healthcare-using-reverse-time-attention-mechanism.pdf%20http://arxiv.org/abs/1608.05745.

Chu, Wei and Zoubin Ghahramani (2005a). "Gaussian Processes for Ordinal Regression". In: *J. Mach. Learn. Res.* 6, pp. 1019–1041. URL: http://dl.acm.org/citation.cfm?id=1046920.1088707.

— (2005b). "Gaussian processes for ordinal regression". In: *Journal of Machine Learning Research* 6.Jul, pp. 1019–1041.

— (2005c). "Preference Learning with Gaussian Processes". In: *Proceedings of the 22Nd International Conference on Machine Learning*. ICML '05. New York, NY, USA: ACM, pp. 137–144. URL: http://doi.acm.org/10.1145/1102351.1102369.

— (2005d). "Preference learning with Gaussian processes". In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pp. 137–144. URL: http://portal.acm.org/citation.cfm?doid=1102351.1102369.

Civil Service World (2017). *Report: Civil servants 'should not fear rise of automation' | Civil Service World*. URL: https://www.civilserviceworld.com/articles/news/report-civil-servants-%E2%80%98should-not-fear-rise-automation%E2%80%99.

Commons, Wikimedia (2019). *File:Rock-paper-scissors.svg — Wikimedia Commons, the free media repository*. URL: https://commons.wikimedia.org/w/index.php?title=File:Rock-paper-scissors.svg&oldid=360388255.

Covington, Martin V. and Kimberly J. Müeller (2001). "Intrinsic Versus Extrinsic Motivation: An Approach/Avoidance Reformulation". en. In: *Educational Psychology Review* 13.2, pp. 157–176.

Cox, Richard T. (1946). "Probability, frequency and reasonable expectation". In: *American journal of physics* 14.1, pp. 1–13.

Crainiceanu, Ciprian M and A Jeffrey Goldsmith (2010). "Bayesian Functional Data Analysis Using WinBUGS." In: *Journal of statistical software* 32.11.

Deming, D. (2015). *The Growing Importance of Social Skills in the Labor Market*. Tech. rep. 21473. NBER Working Paper.

Diamond, John B., Antonia Randolph, and James P. Spillane (2004). "Teachers'Expectations and Sense of Responsibility for Student Learning: The Importance of Race, Class, and Organizational Habitus". en. In: *Anthropology & Education Quarterly* 35.1, pp. 75–98.

D.O.E., U.S. (2017). "College Scorecard Data". In: *College Scorecard Data*. URL: https://collegescorecard.ed.gov/data/.

Doshi-Velez, Finale and Been Kim (2017). "A Roadmap for a Rigorous Science of Interpretability". In: Ml, pp. 1–12. arXiv: 1702.08608. URL: http://arxiv.org/abs/1702.08608.

Doyle, Jon (2004). "Prospects for Preferences". In: *Computational Intelligence* 20.2, pp. 111–136. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.0824-7935.2004.00233.x/abstract.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. URL: http://archive.ics.uci.edu/ml.

Duffy, Mignon, Amy Armenia, and Clare L. Stacey (2015). *Caring on the Clock: The Complexities and Contradictions of Paid Care Work*. en. Google-Books-ID: vdj3BQAAQBAJ. Rutgers University Press.

Duvenaud, David, Hannes Nickisch, and Carl Edward Rasmussen (2011). "Additive Gaussian Processes". In: *Advances in Neural Information Processing Systems 24*, pp. 1–9. arXiv: arXiv:1112.4394v1. URL: http://eprints.pascal-network.org/archive/00008445/.

Education Technology (2017). *Future jobs require upskilling not robots*. Education Technology. URL: https://edtechnology.co.uk/Article/future-jobs-require-upskilling-not-robots/.

Engelbrecht, A. P., I. Cloete, and J. M. Zurada (1995). "Determining the significance of input parameters using sensitivity analysis". In: *International Workshop on Artificial Neural Networks*. Springer, Berlin, Heidelberg, pp. 382–388. URL: http://link.springer.com/10.1007/3-540-59497-3%5C_199.

Fan, Yingying et al. (2014). "Functional Response Additive Model Estimation with Online Virtual Stock Markets". In: pp. 1–31.

Febrero-Bande, Manuel and Manuel Oviedo de la Fuente (2012). "Statistical computing in functional data analysis: the R package fda. usc". In: *Journal of Statistical Software* 51.4, pp. 1–28. URL: http://www.jstatsoft.org/v51/i04/paper.

Fed, U.S. (2018). "Consumer Credit Series Reference". In: *Consumer Credit*. URL: https://www.federalreserve.gov/releases/g19/current/default.htm.

Ferraty, Frédéric, André Mas, and Philippe Vieu (2006). "Advances on nonparametric regression for functional variables". In: arXiv: 0603084 [math]. URL: http://arxiv.org/abs/math/0603084.

Ferraty, Frédéric and Philippe Vieu (2006). *Nonparametric Functional Data Analysis - Theory and Practice*, p. 258.

Ferraty, F. et al. (2013). "Functional projection pursuit regression". In: *Test* 22.2, pp. 293–320.

Frey, Carl Benedikt and Michael A. Osborne (2014). *Agiletown: the relentless march of technology and London's response*. Tech. rep. Deloitte. URL: http://www.deloitte.com/view/en%5C_GB/uk/market-insights/uk-futures/london-futures/index.htm.

Frey, Carl Benedikt and Michael A. Osborne (2017). "The future of employment: how susceptible are jobs to computerisation?" In: *Technological Forecasting and Social Change* 114, pp. 254–280.

Ghahramani, Zoubin (2013). "Bayesian non-parametrics and the probabilistic approach to modelling". In: *Phil. Trans. R. Soc. A* 371.1984, p. 20110553.

Ghavamzadeh, Mohammad, Yaakov Engel, and Michal Valko (2016). "Bayesian Policy Gradient and Actor-Critic Algorithms". In: *Journal of Machine Learning Research* 17.66, pp. 1–53. URL: http://jmlr.org/papers/v17/10-245.html.

Goldsmith, Jeff et al. (2011). "Penalized Functional Regression". In: *Journal of Computational and Graphical Statistics* 20.4, pp. 830–851. arXiv: NIHMS150003.

Goodman, Bryce and Seth Flaxman (2016). "European Union regulations on algorithmic decision-making and a "right to explanation"". In: arXiv: 1606.08813. URL: http://arxiv.org/abs/1606.08813.

Guidotti, Riccardo et al. (2018). "A Survey Of Methods For Explaining Black Box Models". In: arXiv: 1802.01933.

Gunter, Tom et al. (2014). "Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature". In: *NIPS*. arXiv: 1411.0439. URL: http://papers.nips.cc/paper/5483-sampling-for-inference-in-probabilistic-models-with-fast-bayesian-quadrature.

Guyon, Isabelle and André Elisseeff (2003). "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar, pp. 1157–1182.

Hampden-Thompson, Gillian and Judith Bennett (2013). "Science Teaching and Learning Activities and Students'Engagement in Science". In: *International Journal of Science Education* 35.8, pp. 1325–1343.

Handel, M. (2016). "Dynamics of Occupational Change: Implications for the Occupational Requirements Survey". In: . mimeo.

Hara, Satoshi and Kohei Hayashi (2016). "Making Tree Ensembles Interpretable". In: *ICML Workshop on Human Interpretability in Machine Learning*. URL: https://arxiv.org/pdf/1606.05390.pdf.

Hastie, Trevor and Robert Tibshirani (1986). "Generalized Additive Models". In: *Statistical Science* 1.3, pp. 297–310. arXiv: arXiv:1011.1669v3. URL: http://projecteuclid.org/euclid.ss/1177013604.

Hausmann, Ricardo et al. (2014). *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. MIT Press.

Hechtlinger, Yotam (2016). "Interpretation of Prediction Models Using the Input Gradient". In: arXiv: 1611.07634. URL: https://arxiv.org/pdf/1611.07634.pdf.

Institute of Medicine (US) (2008). *Retooling for an Aging America: Building the Health Care Workforce*. Washington (DC): National Academies Press (US).

James, G (2002). "Generalized Linear Models with Functional Predictor Variables". In: *Journal of the Royal Statistical Society, Series B* 64.2, pp. 411–432.

Jaynes, Edwin T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Joachims, Thorsten (2002). "Optimizing Search Engines using Clickthrough Data". In: p. 10.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Press.

Kautz, Tim et al. (2014). *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*. Working Paper 20749. National Bureau of Economic Research.

Kim, Been, Julie Shah, and Finale Doshi-Velez (2015). "Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction". In: *NIPS*. URL: http://people.csail.mit.edu/beenkim/papers/BKim2015NIPS.pdf.

King, Davis E (2009). "Dlib-ml: A Machine Learning Toolkit". In: p. 4.

King, Miriam et al. (2010). "Integrated public use microdata series, current population survey: Version 3.0.[machine-readable database]". In: *Minneapolis: University of Minnesota* 20. URL: http://doi.org/10.18128/D030.V4.0.

Kiureghian, Armen Der and Ove Ditlevsen (2009). "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2, pp. 105–112. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167473008000556.

Klimenko, A. Y. (2015). "Intransitivity in Theory and in the Real World". In: *Entropy* 17.12, pp. 4364–4412. URL: http://arxiv.org/abs/1507.03169.

Koh, Pang Wei and Percy Liang (2017). "Understanding Black-box Predictions via Influence Functions". In: arXiv: 1703.04730. URL: https://arxiv.org/pdf/1703.04730.pdf.

Kumar, Devinder, Alexander Wong, and Graham W Taylor (2017). "Explaining the Unexplained: A CLass-Enhanced Attentive Response (CLEAR) Approach to Understanding Deep Neural Networks". In: arXiv: 1704.04133. URL: https://arxiv.org/pdf/1704.04133.pdf%20http://arxiv.org/abs/1704.04133.

Kuss, Malte and Carl Edward Rasmussen (2006). "Assessing Approximations for Gaussian Process Classification". In: *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, pp. 699–706. URL: http://www.is.tuebingen.mpg.de/fileadmin/user%5C_upload/files/publications/NIPS2005%5C_0163%5C_3530[1].pdf.

Lakkaraju, Himabindu, Stephen H. Bach, and Jure Leskovec (2016). "Interpretable Decision Sets". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, pp. 1675–1684. URL: http://dl.acm.org/citation.cfm?doid=2939672.2939874.

Lipton, Zachary C. (2016). "The Mythos of Model Interpretability". In: arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490.

Lucas, Bill, Guy Claxton, and Ellen Spencer (2013). *Progression in Student Creativity in School*. en. OECD Education Working Papers. Paris: Organisation for Economic Co-operation and Development.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Malakooti, B. (2000). "Ranking and screening multiple criteria alternatives with partial information and use of ordinal and cardinal strength of preferences". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30.3, pp. 355–368.

Martin Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: http://tensorflow.org/.

Matthews, Alexander G. de G., Mark van der Wilk, Tom Nickson, Keisuke. Fujii, et al. (2016). "GPflow: A Gaussian process library using TensorFlow". In: *arXiv preprint 1610.08733*.

Matthews, Alexander G. de G., Mark van der Wilk, Tom Nickson, Keisuke Fujii, et al. (2016). "GPflow: A Gaussian process library using TensorFlow". In: *Journal of Machine Learning Research* 18, pp. 1–6. arXiv: 1610.08733. URL: http://www.jmlr.org/papers/volume18/16-537/16-537.pdf%20http://arxiv.org/abs/1610.08733.

McKinsey Global Institute (2017). "A Future That Works: Automation, Employment and Productivity". In:

McLean, Mathew W et al. (2012). "Functional Generalized Additive Models." In: *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 23.1, pp. 249–269. URL: http://www.ncbi.nlm.nih.gov/pubmed/24729671.

McLean, Mathew W. et al. (2014). "Bayesian Functional Generalized Additive Models with Sparsely Observed Covariates". In: *arXiv*, p. 36. arXiv: 1305.3585. URL: https://arxiv.org/pdf/1305.3585.pdf.

Miller, Tim (2017). "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *arXiv:1706.07269 [cs]*. arXiv: 1706.07269. URL: http://arxiv.org/abs/1706.07269.

MIRON, Marius (2018). *Interpretability in AI and its relation to fairness, transparency, reliability and trust*. JRC Science Hub Communities - European Commission. URL: https://ec.europa.eu/jrc/communities/en/community/humaint/article/interpretability-ai-and-its-relation-fairness-transparency-reliability-and.

Morris, Jeffrey S. (2015). "Functional Regression". In: *Annual Review of Statistics and Its Application* 2.1, pp. 321–359. arXiv: arXiv:1406.4068v1. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-010814-020413.

Müller, Hans Georg and Ulrich Stadtmüller (2005). "Generalized functional linear models". In: *Annals of Statistics* 33.2, pp. 774–805. arXiv: 0505638 [math].

Müller, Hans-Georg and Fang Yao (2008). "Functional Additive Models". In: *Journal of the American Statistical Association* 103.484, pp. 1534–1544.

Muller, H.-G., Yichao Wu, and Fang Yao (2013). "Continuously additive models for nonlinear functional regression". In: *Biometrika* 100.3, pp. 607–622.

Multiple-criteria decision analysis (2020). In: *Wikipedia*. Page Version ID: 937013826. URL: https://en.wikipedia.org/w/index.php?title=Multiple-criteria_decision_analysis&oldid=937013826.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.

Neal, Radford M (1995). "Bayesian Learning for Neural Networks". PhD thesis. University of Toronto.

Occupational Information Network (O*NET) (2017). *O*NET OnLine*. URL: https://www.onetonline.org.

OECD (2016a). "Fostering and Assessing Students'Creativity and Critical Thinking in Higher Education". en. In: *Workshop Summary Report*. Organisation for Economic Co-operation and Development.

— (2016b). "How Good Is Your Job? Measuring and Assessing Job Quality". en. In: *OECD Employment Outlook*. Organisation for Economic Co-operation and Development, pp. 79–139.
— (2017). "PISA 2015 Results (Volume 1): Excellence and Equity". en. In: Organisation for Economic Co-operation and Development Publishing.
O'Hagan, A (1991). "Bayes–hermite quadrature". In: *Journal of Statistical Planning and Inference*. URL: http://www.sciencedirect.com/science/article/pii/037837589190002V.
O'Hagan, A. (1992). *Some Bayesian Numerical Analysis*.
O'Neil, Cathy (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group. URL: http://dl.acm.org/citation.cfm?id=3002861.
Opper, Manfred and Cédric Archambeau (2008). "The Variational Gaussian Approximation Revisited". In: *Neural Computation* 21.3, pp. 786–792. URL: https://doi.org/10.1162/neco.2008.08-07-592.
Orbanz, Peter and Yee Whye Teh (2011). "Bayesian nonparametric models". In: *Encyclopedia of Machine Learning*. Springer, pp. 81–89.
Osborne, Michael (2010). "Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature". In: *Quadrature*, p. 234. URL: http://www.robots.ox.ac.uk/~mosb/public/pdf/136/full%5C_thesis.pdf.
Osborne, Michael a et al. (2012). "Bayesian Quadrature for Ratios". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* 22, pp. 832–840. URL: http://www.jmlr.org/proceedings/papers/v22/osborne12/osborne12.pdf.
Pearson, Karl (1901). "On lines and planes of closest fit to systems of points in space". In: *Philosophical Magazine Series 6* 2.11, pp. 559–572. eprint: http://dx.doi.org/10.1080/14786440109462720. URL: http://dx.doi.org/10.1080/14786440109462720.
Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
Pfingsten, Tobias (2006). "Bayesian Active Learning for Sensitivity Analysis". In: *LNAI* 4212, pp. 354–365. URL: http://www.kyb.tue.mpg.de/fileadmin/user%5C_upload/files/publications/attachments/Pfingsten%5C_2006%5C_Bayesian%5C_Active%5C_Learning%5C_4095%5C%255B0%5C%255D.pdf.
Public Technology (2017). *Don't fear the robots – tech automation to have less impact on public sector jobs, report claims | PublicTechnology.net*. URL: https://www.publictechnology.net/articles/news/don%E2%80%99t-fear-robots-%E2%80%93-tech-automation-have-less-impact-public-sector-jobs-report-claims.
PwC (2017). "Consumer Spending Prospects and the Impact of Automation on Jobs". In: . UK Economic Outlook March.
Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis (2nd Ed)*. New York: Springer.
Rasmussen, Carl Edward (2006a). "Gaussian processes for machine learning". In:
— (2006b). "Gaussian processes in machine learning". In: *International journal of neural systems* 14.2, pp. 69–106. arXiv: 026218253X.

Rasmussen, Carl Edward and Zoubin Ghahramani (2003). "Bayesian Monte Carlo". In: *Advances in Neural Information Processing Systems 15* 15.1, pp. 489–496. URL: http://www.gatsby.ucl.ac.uk.

Rasmussen, C.E. and C. K. Williams (2006). *Gaussian Processes for Machine Learning*. the MIT Press. URL: http://www.gaussianprocess.org/gpml/chapters/.

Rasmussen, CE and CKI Williams (2006). *Gaussian processes for machine learning*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). "Model-Agnostic Interpretability of Machine Learning". In: arXiv: 1606.05386. URL: http://arxiv.org/abs/1606.05386.

— (2016b). ""Why Should I Trust You?"" In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, pp. 1135–1144. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf%20http://dl.acm.org/citation.cfm?doid=2939672.2939778.

Roberts, S et al. (2013). "Gaussian processes for time-series modelling." In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 371.1984, p. 20110550. URL: http://www.ncbi.nlm.nih.gov/pubmed/23277607.

Schunk, Dale H. and Barry J. Zimmerman (2007). "Influencing Children's Self-Efficacy and Self-Regulation of Reading and Writing Through Modeling". In: *Reading & Writing Quarterly* 23.1, pp. 7–25.

Shi, J. Q., B. Wang, R. Murray-Smith, et al. (2007). "Gaussian process functional regression modeling for batch data". In: *Biometrics* 63.3, pp. 714–723.

Shi, J. Q., B. Wang, E. J. Will, et al. (2012). "Mixed-effects Gaussian process functional regression models with application to dose-response curve prediction". In: *Statistics in Medicine* 31.26, pp. 3165–3177.

Shiller, Robert J. (2017). *Understanding Today's Stagnation*. URL: https://www.project-syndicate.org/commentary/secular-stagnation-future-of-work-fears-by-robert-j--shiller-2017-05.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: arXiv: 1312.6034. URL: http://arxiv.org/abs/1312.6034.

Snoek, Jasper, Hugo Larochelle, and Ryan P Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *NIPS*. URL: http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf.

Taylor, Matthew (2017). *Good Work: The Taylor Review of Modern Working Practice*. Tech. rep. An independent review of modern working practices. Department for Business, Energy and Industrial Strategy.

Tes (2017). *Apprenticeships based on 'overly simplistic' view of labour market*. Tes. URL: https://www.tes.com/news/apprenticeships-based-overly-simplistic-view-labour-market.

Tett, G. (2017). *An anthropologist in the boardroom*. URL: https://www.ft.com/content/38e276a2-2487-11e7-a34a-538b4cb30025.

The Bookseller (2017). *Pearson skills report predicts 'one in five' jobs heading for decline | The Bookseller*. URL: https://www.thebookseller.com/news/pearson-skills-report-predicts-only-one-five-jobs-heading-decline-644826.

Titsias, Michalis K and Neil D Lawrence. "Bayesian Gaussian Process Latent Variable Model". In: p. 8.

Toro, Gabriel R. et al. (2010). "Approaches for the efficient probabilistic calculation of surge hazard". In: *Ocean Engineering*. URL: http://www.waveworkshop.org/10thWaves/Papers/Toro%5C_et%5C_al%5C_Wave%5C_Workshop%5C_paper.pdf.

UK Commission for Employment and Skills (2014). *The Future of Work: Jobs and Skills in 2030*. Tech. rep. Evidence Report 84.

U.S. Bureau of Labor Statistics (2010). *2010 SOC User Guide*. URL: https://www.bls.gov/soc/soc%5C_2010%5C_user%5C_guide.pdf.

— (2015). *May 2015 Occupational Employment Statistics*. URL: https://www.bls.gov/oes/2015/may/oes%5C_nat.htm.

— (2016). *Replacement Needs*. URL: https://www.bls.gov/emp/ep%5C_table%5C_110.htm.

Vellido, Alfredo, J D Martin-Guerroro, and P Lisboa (2012). "Making machine learning models interpretable". In: *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. April, pp. 163–172. URL: https://pdfs.semanticscholar.org/ce0b/8b6fca7dc089548cc2e9aaac3bae82bb19da.pdf.

Waegeman, Willem, Bernard De Baets, and Luc Boullart (2008). "ROC analysis in ordinal regression learning". In: *Pattern Recognition Letters* 29.1, pp. 1–9.

Wang, Bo, Tao Chen, and Aiping Xu (2017). "Gaussian process regression with functional covariates and multivariate response". In: *Chemometrics and Intelligent Laboratory Systems* 163.February, pp. 1–6. URL: http://dx.doi.org/10.1016/j.chemolab.2017.02.001.

Wang, Bo and Jian Qing Shi (2014). "Generalized Gaussian Process Regression Model for Non-Gaussian Functional Data". In: *Journal of the American Statistical Association* March, pp. 00–00. arXiv: arXiv:1401.8189v1. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.2014.889021.

Wang, Jane-Ling, Jeng-Min Chiou, and Hans-Georg Müller (2016). "Functional Data Analysis". In: *Annual Review of Statistics and Its Application* 3.1, pp. 257–295. arXiv: 1507.05135.

Weinstein, Rhona S. (2002). *Reaching Higher*. en. Google-Books-ID: B7aP0aa6EswC. Harvard University Press.

Weisstein, Eric W. *Gaussian Quadrature*. Library Catalog: mathworld.wolfram.com Publisher: Wolfram Research, Inc. URL: http://mathworld.wolfram.com/GaussianQuadrature.html.

Yao, Fang and Hans Georg Müller (2010). "Functional quadratic regression". In: *Biometrika* 97.1, pp. 49–64.

Yao, Fang, Hans-Georg Muller, and Jane-ling Wang (2004). "Functional Linear Regression Analysis for Longitudinal Data". In: *The Annals of Statictics* 33.6, pp. 2873–2903.