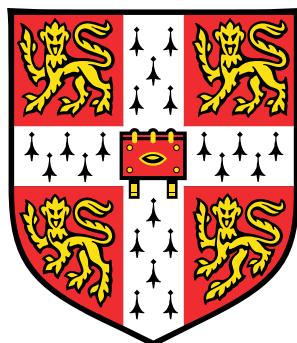


# Automatic Model Construction with Gaussian Processes



**David Kristjanson Duvenaud**

University of Cambridge

This dissertation is submitted for the degree of

*Doctor of Philosophy*

Pembroke College

May 2014

## **Declaration**

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 65,000 words including appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

David Kristjanson Duvenaud  
May 2014

## Acknowledgements

First, I would like to thank my supervisor, Carl Rasmussen, for so much advice and encouragement. It was wonderful working with someone who had spent many years thinking deeply about the business of modeling. Carl was patient while I spent the first few months of my PhD chasing half-baked ideas, and then gently suggested a series of ideas which actually worked.

I'd also like to thank my advisor, Zoubin Ghahramani, for providing much encouragement, support, feedback and advice, for showing me how to solve difficult problems, and for being an example. As well, one of the best things about the lab was that it was constantly populated with interesting visitors, thanks mainly to Zoubin.

Thanks to Michael Osborne, whose thesis was an inspiration early on, and to Roman Garnett for being a sane third party. Sinead Williamson, Peter Orbanz and John Cunningham gave valuable advice when I was still bewildered. Ferenc Huszar and David Knowles made it clear that constantly asking questions was the right way to go.

Tom Dean and Greg Corrado made me feel at home at Google. Philipp Hennig was a mentor to me during my time in Tübingen, and demonstrated an inspiring level of effectiveness. Who else finishes their conference papers with entire days to spare? I'd also like to thank Ryan Adams for hosting me during my visit to Harvard, for our enjoyable collaborations, and for building such an awesome group. Josh Tenenbaum's broad perspective and enthusiasm made the projects we worked on very rewarding. The time I got to spend with Roger Grosse was mind-expanding – he constantly surprised me by pointing out basic unsolved questions about decades-old methods, and had extremely high standards for his own work.

I'd like to thank Andrew McHutchon, Konstantina Palla, Alex Davies, Neil Houlsby, Koa Heaukulani, Miguel Hernández-Lobato, Yue Wu, Ryan Turner, Roger Frigola, Sae Franklin, David Lopez-Paz, Mark van der Wilk and Rich Turner for making the lab feel like a family. I'd like to thank James Lloyd for many endless and enjoyable discussions, and for keeping a level head even during the depths of deadline death-marches. Christian Steinruecken showed me that amazing things are hidden all around.

Thanks to Mel for reminding me that I needed to write a thesis, and for supporting me during the final push.

My graduate study was supported by the National Sciences and Engineering Research Council of Canada, the Cambridge Commonwealth Trust, Pembroke College, a grant from the Engineering and Physical Sciences Research Council, and a grant from Google.

# Abstract

This thesis develops a method for automatically constructing, illustrating and describing a large class of models, useful for forecasting and finding structure in domains such as time series, geological formations, and physical dynamics. These models, based on Gaussian processes, can capture many types of statistical structure, such as periodicity, changepoints, additivity, and symmetries. Such structure can be encoded through a *kernel*, which has historically been chosen by hand by experts. We show how to automate this task, creating a system which explores a large space of models and reports the structures discovered.

The introductory chapters show how to express many types of structure through kernels, and how adding and multiplying different kernels combines their properties. Among several examples, we show how composite kernels can produce priors over topological manifolds such as cylinders, toruses, and Möbius strips, as well as their higher-dimensional analogues.

To automatically search over an open-ended space of models, we define a simple grammar over kernels, a search criterion (marginal likelihood), and a breadth-first search procedure. Combining these, we present a procedure which takes in a dataset and outputs an automatically-constructed model, along with a detailed report with graphs and automatically generated text illustrating the qualitatively different, and sometimes novel, types of structure discovered in that dataset. This system automates parts of the model-building and analysis currently performed by expert statisticians.

This thesis also explores several extensions to Gaussian process models. First, building on earlier work relating Gaussian processes and neural nets, we explore the natural extensions of these models to *deep kernels* and *deep Gaussian processes*. Second, we examine the model class consisting of the sum of functions of all possible combinations of input variables. We show a close connection between this model class and the recently-developed regularization method of *dropout*. Third, we combine Gaussian processes with the Dirichlet process to produce the *warped mixture model* – a Bayesian clustering model with nonparametric cluster shapes, and a corresponding latent space in which

each cluster has an interpretable parametric form.

# Contents

<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Notation</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Gaussian Process models . . . . .	1
1.1.1 Model selection . . . . .	3
1.1.2 Prediction . . . . .	4
1.1.3 Useful properties of Gaussian processes . . . . .	4
1.1.4 Limitations of Gaussian processes . . . . .	5
1.2 Outline and contributions of thesis . . . . .	6
<b>2 Expressing Structure with Kernels</b>	<b>8</b>
2.1 Definition . . . . .	8
2.2 A few basic kernels . . . . .	9
2.3 Combining kernels . . . . .	10
2.3.1 Notation . . . . .	10
2.3.2 Combining properties through multiplication . . . . .	11
2.3.3 Building flexible multidimensional models . . . . .	12
2.4 Modeling sums of functions . . . . .	13
2.4.1 Additivity across multiple dimensions . . . . .	14
2.4.2 Long-range extrapolation through additivity . . . . .	15
2.4.3 Example: an additive model of concrete strength . . . . .	16
2.4.4 Posterior variance of additive components . . . . .	16
2.5 Changepoints . . . . .	18
2.5.1 Multiplication by a known function . . . . .	20

2.6	Feature representation of kernels . . . . .	21
2.6.1	Relation to linear regression . . . . .	21
2.6.2	Feature-space view of combining kernels . . . . .	21
2.7	Expressing symmetries and invariants . . . . .	22
2.7.1	Three recipes for invariant priors . . . . .	23
2.7.2	Example: periodicity . . . . .	25
2.7.3	Example: symmetry about zero . . . . .	25
2.7.4	Example: translation invariance in images . . . . .	26
2.8	Generating topological manifolds . . . . .	27
2.8.1	Möbius strips . . . . .	28
2.9	Kernels on categorical variables . . . . .	29
2.10	Building a kernel in practice . . . . .	29
<b>3</b>	<b>Automatic Model Construction</b>	<b>30</b>
3.1	Ingredients of an automatic statistician . . . . .	31
3.2	A language of regression models . . . . .	32
3.3	A model search procedure . . . . .	33
3.4	A model comparison procedure . . . . .	36
3.5	A model description procedure . . . . .	36
3.6	Structure discovery in time series . . . . .	37
3.6.1	Mauna Loa atmospheric CO <sub>2</sub> . . . . .	37
3.6.2	Airline passenger counts . . . . .	38
3.7	Related work . . . . .	39
3.8	Experiments . . . . .	42
3.8.1	Interpretability versus accuracy . . . . .	42
3.8.2	Predictive accuracy on time series . . . . .	43
3.8.3	High-dimensional prediction . . . . .	44
3.8.4	Structure recovery on synthetic data . . . . .	44
3.9	Discussion . . . . .	46
<b>4</b>	<b>Automatic Model Description</b>	<b>47</b>
4.1	Generating descriptions of composite kernels . . . . .	48
4.1.1	Simplification rules . . . . .	48
4.1.2	Describing each part of a product of kernels . . . . .	49
4.1.3	Combining descriptions into noun phrases . . . . .	50
4.1.4	Worked example . . . . .	51

4.2	Example descriptions . . . . .	52
4.2.1	Summarizing 400 years of solar activity . . . . .	52
4.2.2	Describing changing noise levels . . . . .	54
4.3	Related work . . . . .	56
4.4	Limitations of this approach . . . . .	56
4.5	Conclusions . . . . .	57
<b>5</b>	<b>Deep Gaussian Processes</b>	<b>58</b>
5.1	Definition and relation to neural networks . . . . .	59
5.1.1	Single-layer models . . . . .	59
5.1.2	Multiple hidden layers . . . . .	60
5.1.3	Two network architectures equivalent to deep GPs . . . . .	61
5.2	Characterizing deep Gaussian process priors . . . . .	63
5.2.1	One-dimensional asymptotics . . . . .	63
5.2.2	Distribution of the Jacobian . . . . .	65
5.3	Formalizing a pathology . . . . .	66
5.4	Fixing the pathology . . . . .	70
5.5	Deep kernels . . . . .	71
5.5.1	Infinitely deep kernels . . . . .	74
5.5.2	When are deep kernels useful models? . . . . .	75
5.6	Related work . . . . .	76
5.7	Conclusions . . . . .	78
<b>6</b>	<b>Additive Gaussian Processes</b>	<b>79</b>
6.1	Different types of multivariate additive structure . . . . .	80
6.2	Defining additive kernels . . . . .	81
6.2.1	Weighting different orders of interaction . . . . .	81
6.2.2	Efficiently evaluating additive kernels . . . . .	82
6.3	Additive models allow non-local interactions . . . . .	84
6.4	Dropout in Gaussian processes . . . . .	85
6.4.1	Dropout on hidden layers . . . . .	85
6.4.2	Dropout on inputs gives additive covariance . . . . .	86
6.5	Related work . . . . .	87
6.6	Regression and classification experiments . . . . .	90
6.6.1	Datasets . . . . .	91
6.6.2	Results . . . . .	91

6.7	Conclusions . . . . .	93
<b>7</b>	<b>Warped Mixture Models</b>	<b>95</b>
7.1	The Gaussian process latent variable model . . . . .	96
7.2	The infinite warped mixture model . . . . .	97
7.3	Inference . . . . .	99
7.4	Related work . . . . .	100
7.5	Experimental results . . . . .	102
7.5.1	Synthetic datasets . . . . .	102
7.5.2	Clustering face images . . . . .	103
7.5.3	Density estimation . . . . .	104
7.5.4	Mixing . . . . .	105
7.5.5	Visualization . . . . .	105
7.5.6	Clustering performance . . . . .	106
7.5.7	Density estimation . . . . .	107
7.6	Conclusions . . . . .	108
7.7	Future work . . . . .	108
<b>8</b>	<b>Bayesian Estimation of Integrals</b>	<b>110</b>
8.1	Log-Gaussian Process Models . . . . .	110
8.2	Linearization . . . . .	111
8.2.1	Mean of $\hat{Z}$ . . . . .	114
8.2.2	Variance of $\hat{Z}$ . . . . .	116
8.2.3	Marginal Variance . . . . .	117
8.3	Nonparametric Inference via Bayesian Quadrature . . . . .	117
8.4	Introduction . . . . .	119
8.5	Bayesian Quadrature . . . . .	119
8.5.1	BQ as an inference method . . . . .	121
8.6	Guidelines for Constructing a Kernel . . . . .	122
8.6.1	Convergence . . . . .	122
8.7	Inference in the Indian Buffet Process . . . . .	123
8.7.1	A kernel between binary matrices . . . . .	124
8.7.2	Computing $z_n$ and the prior variance . . . . .	125
8.8	Infinite Mixture Models . . . . .	125
8.8.1	A Kernel Between Multinomial Distributions . . . . .	126
8.8.2	Infinite Mixture of Gaussians . . . . .	127

8.9	Related Work . . . . .	127
8.10	Experiments . . . . .	127
8.10.1	IBP Experiments . . . . .	128
8.10.2	DP Mixture Experiments . . . . .	128
8.11	Discussion . . . . .	129
8.11.1	Appropriateness of the GP prior . . . . .	129
8.12	Conclusions . . . . .	130
8.13	Gaussian Process Posteriors . . . . .	130
8.13.1	Marginal Mean Function . . . . .	132
8.13.2	Marginal Variance Function . . . . .	133
8.14	Objective Function . . . . .	137
<b>9</b>	<b>Discussion</b>	<b>138</b>
9.1	Structured versus unstructured models . . . . .	138
9.2	Approaches to automating model construction . . . . .	140
9.3	Summary of contributions . . . . .	140
<b>Appendix A</b>	<b>Gaussian conditionals</b>	<b>142</b>
<b>Appendix B</b>	<b>Kernel definitions</b>	<b>143</b>
<b>Appendix C</b>	<b>Search operators</b>	<b>145</b>
<b>Appendix D</b>	<b>Example automatically generated report</b>	<b>147</b>
<b>Appendix E</b>	<b>Inference in the warped mixture model</b>	<b>155</b>
<b>References</b>		<b>160</b>

# List of Figures

1.1	One-dimensional Gaussian process posterior . . . . .	2
2.1	Examples of structures expressible by some basic kernels . . . . .	9
2.2	Examples of structures expressible by multiplying kernels . . . . .	11
2.3	A product of squared-exponential kernels across different dimensions . . . . .	12
2.4	Examples of structures expressible by adding kernels . . . . .	13
2.5	Additive kernels correspond to additive functions . . . . .	14
2.6	Long-range inference in functions with additive structure . . . . .	15
2.7	Decomposition of posterior into interpretable one-dimensional functions .	17
2.8	Visualizing posterior correlations between components . . . . .	19
2.9	Draws from changepoint priors . . . . .	20
2.10	Three ways to introduce symmetry . . . . .	24
2.11	Generating 2D manifolds with different topological structures . . . . .	27
2.12	Generating Möbius strips . . . . .	28
3.1	Another set of basic kernels . . . . .	32
3.2	A search tree over kernels . . . . .	34
3.3	Progression of models as the search depth increases . . . . .	35
3.4	Decomposition of the Mauna Loa model . . . . .	38
3.5	Decomposition of airline dataset model . . . . .	39
3.6	Extrapolation error of all methods on 13 time-series datasets . . . . .	44
4.1	Solar irradiance dataset . . . . .	53
4.2	Automatically-generated descriptions of the solar irradiance data set . . .	53
4.3	A learned component corresponding to the Maunder minimum . . . . .	53
4.4	ABCD isolating the part of the signal explained by a slowly-varying trend	54
4.5	Automatic description of solar cycle . . . . .	54
4.6	Short descriptions of the four components of the airline model . . . . .	55

4.7	Describing non-stationary periodicity in the airline data . . . . .	55
4.8	Modeling time-changing variance in the airline dataset . . . . .	55
5.1	Neural network architectures giving rise to GPs . . . . .	61
5.2	Neural network architectures giving rise to deep GPs . . . . .	62
5.3	A one-dimensional draw from a deep GP prior . . . . .	63
5.4	Desirable properties of representations of manifolds . . . . .	66
5.5	Distribution of singular values of the Jacobian of a deep GP . . . . .	67
5.6	Points warped by a draw from a deep GP . . . . .	68
5.7	Visualization of a feature map drawn from a deep GP . . . . .	69
5.8	Two different architectures for deep neural networks . . . . .	70
5.9	A draw from a 1D deep GP prior with each layer connected to the input	71
5.10	Points warped by a draw from an input-connected deep GP . . . . .	72
5.11	Distribution of singular values of an input-connected deep GP . . . . .	72
5.12	Feature mapping of an input-connected deep GP . . . . .	73
5.13	Infinitely deep kernels . . . . .	75
6.1	Isocontours of additive kernels in 3 dimensions . . . . .	84
6.2	A comparison of different additive model classes . . . . .	89
7.1	One-dimensional Gaussian process latent variable model . . . . .	96
7.2	Two-dimensional Gaussian process latent variable model . . . . .	97
7.3	A draw from the infinite warped mixture model prior . . . . .	98
7.4	Recovering clusters on synthetic data . . . . .	102
7.5	Latent clusters of face images . . . . .	103
7.6	Comparing density estimates of between the GP-LVM and the iWMM . .	104
7.7	A visualization of a sampler for the iWMM . . . . .	105
7.8	Comparison of latent coordinate estimates . . . . .	106
8.1	Comparing the differing characteristics of GP ans log-GP models . . . .	112
8.2	Comparing the differing characteristics of GP ans log-GP models . . . .	118
8.3	An illustration of Bayesian quadrature . . . . .	120
8.4	Computing marginal likelihoods with BQ . . . . .	128
8.5	Computing marginal likelihoods with BQ . . . . .	129
8.6	A 2D GP posterior, and its 1D marginals . . . . .	131

# List of Tables

3.1	Common regression models expressible in the kernel language . . . . .	33
3.2	Kernels recovered on synthetic data . . . . .	45
4.1	Descriptions of the effect of each kernel, written as a post-modifier . . . . .	50
4.2	Noun phrase descriptions of each type of kernel . . . . .	50
6.1	Relative variance contributed by each order of the additive model . . . . .	82
6.2	Regression Dataset Statistics . . . . .	91
6.3	Classification Dataset Statistics . . . . .	91
6.4	Comparison of predictive error on regression problems . . . . .	92
6.5	Comparison of predictive likelihood on regression problems . . . . .	92
6.6	Comparison of predictive error on classification problems . . . . .	92
6.7	Comparison of predictive likelihood on classification problems . . . . .	92
7.1	Datasets used for evaluation of the iWMM . . . . .	106
7.2	Clustering performance comparison . . . . .	107
7.3	Predictive likelihood comparison . . . . .	107

# Notation

Unbolded  $x$  represents a single number, while  $\mathbf{x}$  represents a vector, and  $\mathbf{X}$  represents a matrix. A individual element of a vector will be denoted with a subscript and without boldface. For example, the  $i$ th element of a vector  $\mathbf{x}$  is denoted as  $x_i$ . A bold lower-case number with an index such as  $\mathbf{x}_j$  represents a particular row of matrix  $\mathbf{X}$ .

Symbol	Description
$\mathcal{O}(\cdot)$	The big-O asymptotic complexity of an algorithm.
$A \otimes B$	The Kronecker product of matrices $A$ and $B$ .
SE	The squared-exponential kernel, also known as the radial-basis function (RBF), or Gaussian kernel.
RQ	The rational-quadratic kernel.
Per	The periodic kernel.
Lin	The linear kernel.
WN	The white-noise kernel.
C	The constant kernel.
$\sigma$	The changepoint kernel, $\sigma(x, x') = \sigma(x)\sigma(x')$ , where $\sigma(x)$ is a sigmoidal function such as the logistic function.
$k_a + k_b$	Addition of kernels, shorthand for $k_a(\mathbf{x}, \mathbf{x}') + k_b(\mathbf{x}, \mathbf{x}')$
$k_a \times k_b$	Multiplication of kernels, shorthand for $k_a(\mathbf{x}, \mathbf{x}') \times k_b(\mathbf{x}, \mathbf{x}')$
$k(\mathbf{X}, \mathbf{X})$	The Gram matrix, whose $i, j$ th element is $k(\mathbf{x}_i, \mathbf{x}_j)$ .
$\mathbf{K}$	Shorthand for the Gram matrix $k(\mathbf{X}, \mathbf{X})$
$\mathbf{f}(\mathbf{X})$	A vector of function values, whose $i$ th element is given by $f(\mathbf{x}_i)$ .
$\text{vec}(\mathbf{X})$	The vectorization operator, which concatenates each column of $\mathbf{X}$ into a column vector.
$\text{mod}(i, j)$	The modulo operator: the remainder after dividing $i$ by $j$ .

# Chapter 1

## Introduction

“All models are wrong, but yours are stupid too.”

@ML\_Hipster (2013)

Prediction, extrapolation, and induction are all examples of learning a function from data. There are many ways to learn functions, but one particularly elegant way is by *inference*. Inference procedures set up a group of hypotheses – a *model*, then weight those hypotheses based on how well their predictions match the data. Keeping around all the hypotheses that match the data helps guard against over-fitting. We can also compare different models to find what sorts of structure are present in a dataset.

To be able to learn a wide variety of types of structure, we would like to have an expressive language of models of functions. We would like to be able to represent simple kinds of functions, such as linear functions or polynomials. We would also like to have models of arbitrarily complex functions, specified in terms of high-level properties such as how smooth they are, whether they repeat over time, or which symmetries they have.

This thesis will show how to build such a language using Gaussian processes (GPs), a tractable set of models of very different types of functions. This chapter will introduce the basic properties of GPs. The next chapter will describe many of the types of functions which we know how to model using GPs.

### 1.1 Gaussian Process models

Gaussian processes are a simple and general class of models of functions. To be precise, a GP is any distribution over functions such that any finite subset of function values  $f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_N)$  have a joint Gaussian distribution (Rasmussen and Williams,

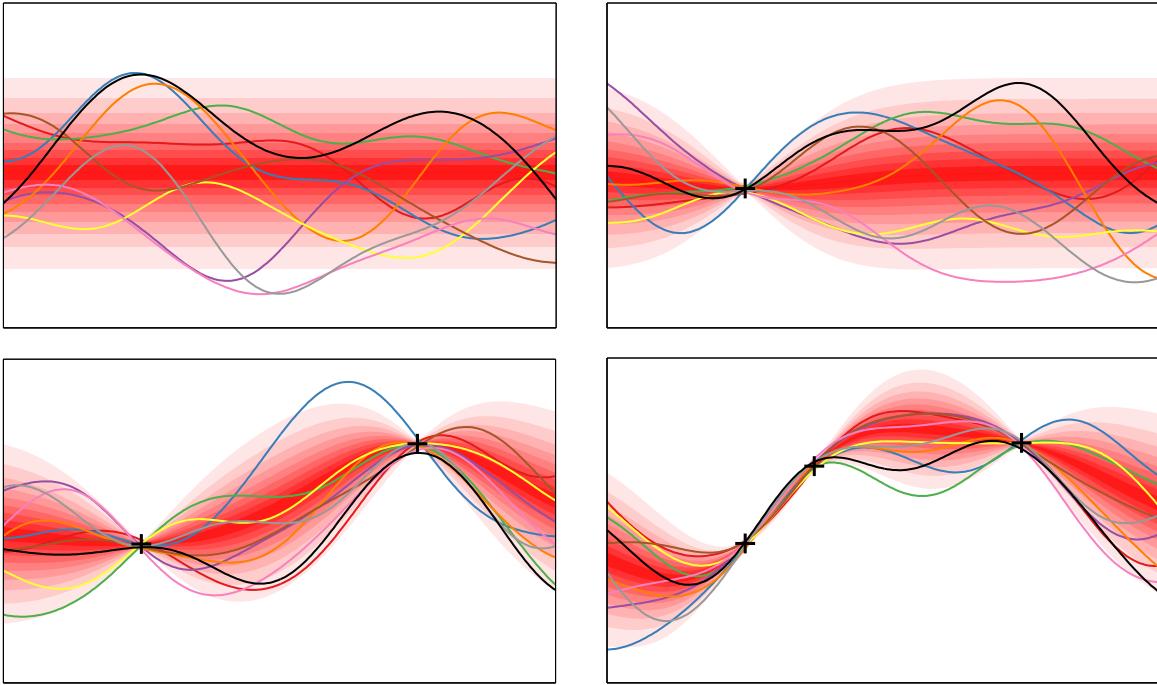


Figure 1.1 A visual representation of a Gaussian process posterior over a one-dimensional function. Different shades of red correspond to deciles of the predictive density at each input location. Coloured lines show samples from the process – examples of some of the hypotheses included in the model. *Top left:* A GP not conditioned on any datapoints. *Remaining plots:* The posterior after conditioning on different amounts of data. All plots have the same axes.

2006). A GP model, before conditioning on data, is completely specified by its mean function,

$$\mathbb{E}[f(\mathbf{x})] = \mu(\mathbf{x}) \quad (1.1)$$

and its covariance function, also called the *kernel*:

$$\text{Cov}(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}') \quad (1.2)$$

It is common practice to assume that the mean function is simply zero everywhere, since uncertainty about the mean function can be taken into account by adding an extra term to the kernel.

After accounting for the mean, the kind of structure which can be captured by a GP model is entirely determined by its kernel. The kernel determines how the model

generalizes, or extrapolates to new data.

There are many possible choices of covariance function, and we can specify a wide range of models just by specifying the kernel of a GP. For example, linear regression, splines, and Kalman filters are all examples of GPs with particular kernels. However, these are just a few familiar examples out of a wide range of possibilities. One of the main difficulties in using GPs is constructing a kernel which represents the particular structure present in the data being modelled.

### 1.1.1 Model selection

The crucial property of GPs that allows us to automatically construct models is that we can compute the *marginal likelihood* of a dataset given a particular model, also known as the *evidence* (MacKay, 1992). The marginal likelihood allows us to compare models and automatically discover the appropriate amount of detail to use, due to Bayesian Occam's razor (MacKay, 2003; Rasmussen and Ghahramani, 2001).

Choosing a kernel, or kernel parameters, by maximizing the marginal likelihood will typically result in selecting the *least* flexible model which still captures all the structure in the data. For example, if a kernel has a parameter which controls the smoothness of the functions it models, the maximum-likelihood estimate of that parameter will usually correspond to the smoothest family of functions which still go through all the observed function values.

Here is the marginal likelihood under a GP prior of observing a set of function values  $[f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_N)] := \mathbf{f}(\mathbf{X})$  at locations  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^\top := \mathbf{X}$ :

$$\begin{aligned} p(\mathbf{f}(\mathbf{X})|\mathbf{X}, \mu(\cdot), k(\cdot, \cdot)) &= \mathcal{N}(\mathbf{f}(\mathbf{X})|\mu(\mathbf{X}), k(\mathbf{X}, \mathbf{X})) \\ &= (2\pi)^{-\frac{N}{2}} \times \underbrace{|k(\mathbf{X}, \mathbf{X})|^{-\frac{1}{2}}}_{\text{discourages flexibility}} \\ &\quad \times \underbrace{\exp \left\{ -\frac{1}{2} (\mathbf{f}(\mathbf{X}) - \boldsymbol{\mu}(\mathbf{X}))^\top k(\mathbf{X}, \mathbf{X})^{-1} (\mathbf{f}(\mathbf{X}) - \boldsymbol{\mu}(\mathbf{X})) \right\}}_{\text{encourages fit with data}} \end{aligned} \tag{1.3}$$

This multivariate Gaussian density is referred to as the *marginal likelihood* because it implicitly integrates (marginalizes) over all possible functions values  $\mathbf{f}(\bar{\mathbf{X}})$ , where  $\bar{\mathbf{X}}$  is the set of all locations where we have not observed the function.

### 1.1.2 Prediction

We can ask the model which function values are likely to occur at any location, given the observations seen so far. By the formula for Gaussian conditionals (given in appendix A), the predictive distribution of a function value  $f(\mathbf{x}^*)$  at a test point  $\mathbf{x}^*$  has the form:

$$p(f(\mathbf{x}^*) | \mathbf{f}(\mathbf{X}), \mathbf{X}, \mu(\cdot), k(\cdot, \cdot)) = \mathcal{N}\left(f(\mathbf{x}^*) | \underbrace{\mu(\mathbf{x}^*) + k(\mathbf{x}^*, \mathbf{X})k(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{f}(\mathbf{X}) - \mu(\mathbf{X}))}_{\text{predictive mean goes through observations}}, \underbrace{k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})k(\mathbf{X}, \mathbf{X})^{-1}k(\mathbf{X}, \mathbf{x}^*)}_{\text{predictive variance shrinks given more data}}\right) \quad (1.4)$$

These expressions may look complex, but only require a few matrix operations to evaluate.

Sampling a function from a GP is also straightforward: a sample from a GP at a finite set of locations is just a single sample from a single multivariate Gaussian distribution, given by equation (1.4). Figure 1.1 shows prior and posterior samples from a GP, as well as contours of the predictive density.

Our use of probabilities does not mean that we are assuming the function being learned is stochastic or random in any way; it is simply a consistent method of keeping track of our uncertainty.

### 1.1.3 Useful properties of Gaussian processes

There are several reasons why GPs in particular are well-suited for building a language of regression models:

- **Analytic inference.** Given a kernel function and some observations, the predictive posterior distribution can be computed exactly in closed form. This is a rare property for nonparametric models to have.
- **Expressivity.** Through the choice of covariance function, we can express a very wide range of modeling assumptions.
- **Integration over hypotheses.** The fact that a GP posterior, given a fixed kernel, lets us integrate exactly over a wide range of hypotheses means that overfitting is less of an issue than in comparable model classes, such as neural networks.

Compared to neural networks, relatively few parameters need to be estimated, which lessens the need for the complex optimization or regularization schemes.

- **Model Selection.** A side benefit of being able to integrate over all hypotheses is that we can compute the marginal likelihood of the data given a model. This gives us a principled way of comparing different models.
- **Closed-form predictive distribution.** The predictive distribution of a GP at a set of test points is simply a multivariate Gaussian distribution. This means that GPs can easily be composed with other models or decision procedures.
- **Easy to analyze.** It may seem unsatisfying to restrict ourselves to a limited model class, as opposed to trying to do inference in the set of all computable functions. However, simple models can be used as well-understood building blocks for constructing more interesting models.

For example, consider linear models. Although they form an extremely limited model class, they are simple, easy to analyze, and easy to incorporate into other models or procedures. Gaussian processes can be seen as an extension of linear models which retain these attractive properties (Rasmussen and Williams, 2006).

#### 1.1.4 Limitations of Gaussian processes

There are, unfortunately, several issues which make usage of GPs sometimes difficult:

- **Slow inference.** Computing the matrix inverse in equations (1.3) and (1.4) takes  $\mathcal{O}(N^3)$  time, making inference prohibitively slow for more than a few thousand datapoints. However, this problem can be addressed by approximate inference schemes (Hensman et al., 2013; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006). Most GP software packages implement several of these methods (Hensman et al., 2014b; Rasmussen and Nickisch, 2010; Vanhatalo et al., 2013).
- **Light tails of the predictive distribution.** The predictive distribution of a standard GP model is Gaussian. In order to be robust to outliers, or to perform classification, we may wish to use non-Gaussian noise models. Fortunately, mature software packages exist which can automatically perform approximate inference for a wide variety of non-Gaussian likelihoods.

- **The need to choose a kernel.** In practice, the extreme flexibility of GP models means that we are also faced with the difficult task of choosing a kernel. In fact, choosing a useful kernel is equivalent to the problem of learning a good representation of the input. Kernel parameters are usually set automatically by maximizing the marginal likelihood. However, until recently, human experts were still required to choose the parametric form of the kernel from a small set of standard kernels. Chapter 3 will show how the entire construction of kernels can be automated.

## 1.2 Outline and contributions of thesis

The main contribution of this thesis is to show how to automate the discovery and explanation of structure in functions, simply by searching an open-ended language of regression models. This thesis also includes a set of related results showing how Gaussian processes can be extended, or composed with other models.

Chapter 2 is a tutorial showing how to build a wide variety of structured models of functions by constructing appropriate covariance functions. We will also show how GPs can produce nonparametric models of manifolds with diverse topological structures, such as cylinders, toruses and Möbius strips.

Chapter 3 shows how to search over a general, open-ended language of models, built by composing together different kernels. Since we can evaluate each model by the marginal likelihood, we can automatically construct custom models for each dataset by a breadth-first search. The nature of GPs allow the resulting models to be visualized by decomposing them into diverse, interpretable components, each capturing a different type of structure. Capturing this high-level structure sometimes even allows us to extrapolate beyond the range of the data.

One benefit of using a compositional model class is that the resulting models are interpretable. Chapter 4 demonstrates a system which automatically describes the structure implied by a given kernel on a given dataset, generating reports with graphs and English-language text describing the resulting model. We will show several automatic analyses of time-series. Combined with the automatic model search developed in chapter 3, this system represents the beginnings of an “automatic statistician”.

Chapter 5 analyzes deep neural network models by characterizing the prior over functions obtained by composing GP priors to form *deep Gaussian processes*. We show that, as the number of layers increase, the amount of information retained about the

original input diminishes to a single degree of freedom. A simple change to the network architecture fixes this pathology. We relate these models to neural networks, and as a side effect derive several forms of *infinitely deep kernels*.

Chapter 6 examines a more limited, but much faster way of discovering structure using GPs. Specifying a kernel with many different types of structure, we use kernel parameters to discard whichever types of structure are *not* found in the current dataset. The model class we examine is called *additive Gaussian processes*, a model summing over exponentially-many GPs, each depending on a different subset of the input variables. We give a polynomial-time inference algorithm for this model class, and relate it to other model classes. For example, additive GPs are shown to have the same covariance as a GP that uses *dropout*, a recently discovered regularization technique for neural networks.

Chapter 7 develops a Bayesian clustering model in which the clusters have nonparametric shapes, called the infinite warped mixture model. The density manifolds learned by this model follow the contours of the data density, and have interpretable, parametric forms in the latent space. The marginal likelihood lets us infer the effective dimension and shape of each cluster separately, as well as the number of clusters.

# Chapter 2

## Expressing Structure with Kernels

This chapter shows how to use kernels to build models of functions with many different kinds of structure: additivity, symmetry, periodicity, interactions between variables, and changepoints. We will also show several ways to encode group invariants into kernels. Combining a few simple kernels through addition and multiplication will give us a rich, open-ended language of models.

The properties of kernels discussed in this chapter are mostly known in the literature. The original contribution of this chapter is to gather them into a coherent whole and to offer a tutorial showing the implications of different kernel choices, and some of the structures which can be obtained by combining them. To the best of our knowledge, the different types of structure that could be obtained through sums and products of kernels had not been systematically explored before [Duvenaud et al. \(2013\)](#).

### 2.1 Definition

A kernel (also called a covariance function, kernel function, or covariance kernel), is a positive-definite function of two inputs  $\mathbf{x}, \mathbf{x}'$ . In this chapter,  $\mathbf{x}$  and  $\mathbf{x}'$  are usually vectors in a Euclidean space, but kernels can also be defined on graphs, images, discrete or categorical inputs, or even text.

Gaussian process models use a kernel to define the prior covariance between any two function values:

$$\text{Cov}(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}') \quad (2.1)$$

Colloquially, kernels are often said to specify the similarity between two objects. This is

slightly misleading in this context, since what is actually being specified is the similarity between two values of a *function* evaluated on each object. The kernel specifies which functions are likely under the GP prior, which in turn determines the generalization properties of the model.

## 2.2 A few basic kernels

To begin understanding the types of structures expressible by GPs, we will start by briefly examining the priors on functions encoded by some commonly used kernels: the squared-exponential (SE), periodic (Per), and linear (Lin) kernels. These kernels are defined in figure 2.1.

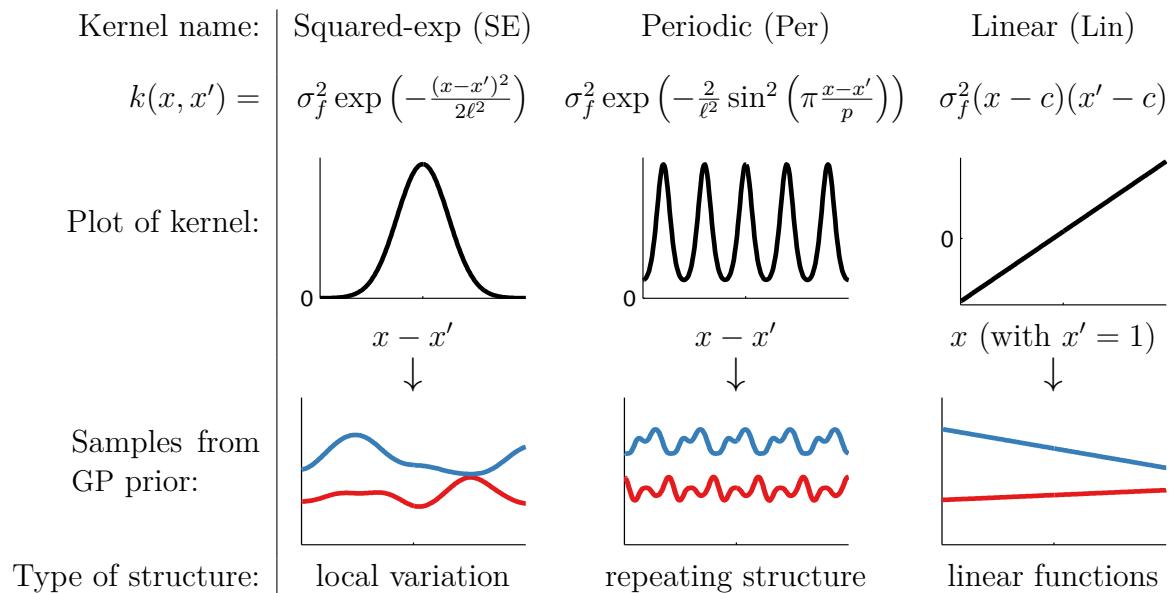


Figure 2.1 Examples of structures expressible by some basic kernels.

Each covariance function corresponds to a different set of assumptions made about the function we wish to model. For example, using a squared-exp (SE) kernel implies that the function we are modeling has infinitely many derivatives. There exist many variants of “local” kernels similar to the SE kernel, each encoding slightly different assumptions about the smoothness of the function being modeled.

**Kernel parameters** Each kernel has a number of parameters which specify the precise shape of the covariance function. These are sometimes referred to as *hyper-parameters*,

since they can be viewed as specifying a distribution over function parameters, instead of being parameters which specify a function directly.

**Stationary and Non-stationary** The SE and Per kernels are *stationary*, meaning that their value only depends on the difference  $x - x'$ . This implies that the probability of observing a particular dataset remains the same even if we move all the  $\mathbf{x}$  values over by some amount. In contrast, the linear kernel Lin is non-stationary, meaning that the corresponding GP model will produce different answers if the data is moved around while the kernel parameters are kept fixed.

## 2.3 Combining kernels

What if the kind of structure we need is not expressed by any existing kernel? For many types of structure, it is possible to build a “made to order” kernel with the desired properties. The next few sections of this chapter will explore ways in which kernels can be combined to create new ones with different properties. This will allow us to include as much high-level structure as necessary into our models.

### 2.3.1 Notation

Below, we will focus on two ways of combining kernels: addition and multiplication. We will often write these operations in shorthand, without arguments:

$$k_a + k_b = k_a(\mathbf{x}, \mathbf{x}') + k_b(\mathbf{x}, \mathbf{x}')$$
 (2.2)

$$k_a \times k_b = k_a(\mathbf{x}, \mathbf{x}') \times k_b(\mathbf{x}, \mathbf{x}')$$
 (2.3)

All of the basic kernels we considered in section 2.2 are one-dimensional, but kernels over multidimensional inputs can be constructed by adding and multiplying between kernels on different dimensions. The dimension on which a kernel operates is denoted by a subscripted integer. For example,  $SE_2$  represents an SE kernel over the second dimension of  $\mathbf{x}$ . To remove clutter, we will usually refer to kernels without specifying their parameters.

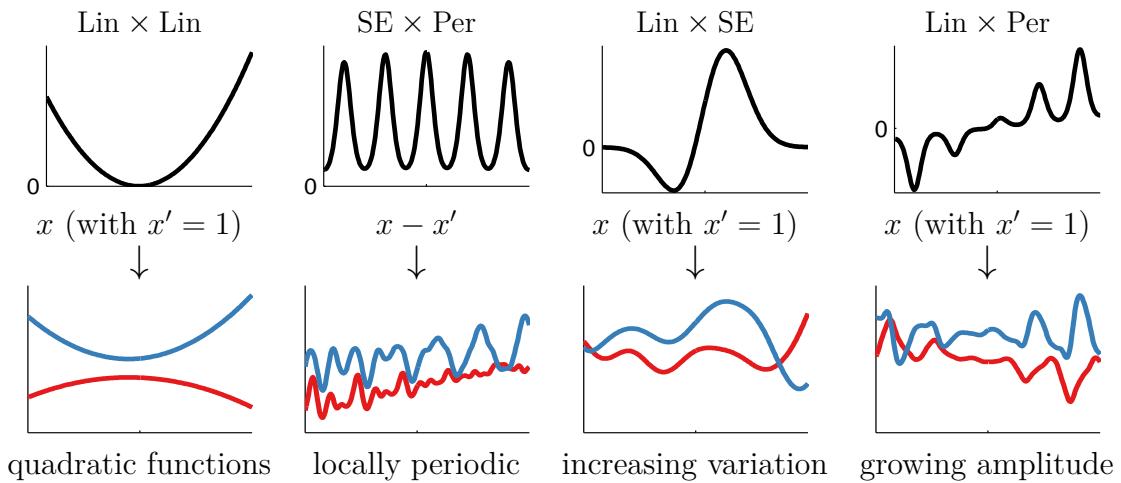


Figure 2.2 Examples of one-dimensional structures expressible by multiplying kernels. Plots have same meaning as in figure 2.1.

### 2.3.2 Combining properties through multiplication

Multiplying two positive-definite kernels together always results in another positive-definite kernel. But what properties do these new kernels have? Figure 2.2 shows some more interesting kernels that one can obtain by multiplying two basic kernels together.

Working with kernels rather than the parametric form of the function itself allows us to express high-level properties of functions that do not necessarily have a simple parametric form. Here, we discuss a few examples:

- **Polynomial Regression.** By multiplying together  $T$  linear kernels, we obtain a prior on polynomials of degree  $T$ . The first column of figure 2.2 shows a quadratic kernel.
- **Locally Periodic Functions.** In univariate data, multiplying a kernel by SE gives a way of converting global structure to local structure. For example, Per corresponds to exactly periodic structure, whereas Per  $\times$  SE corresponds to locally periodic structure, as shown in the second column of figure 2.2.
- **Functions with Growing Amplitude.** Multiplying by a linear kernel means that the marginal standard deviation of the function being modeled will grow linearly away from the origin. The third and fourth columns of figure 2.2 show two examples.

We can multiply any number of kernels together in this way to produce kernels combining several high-level properties. For example, the kernel  $\text{SE} \times \text{Lin} \times \text{Per}$  is a prior on functions which are locally periodic with linearly growing amplitude. We will see examples of real datasets with this kind of structure in chapter 3.

### 2.3.3 Building flexible multidimensional models

We can build flexible models of functions of more than one input simply by multiplying kernels between the different inputs. For example, a product of SE kernels over different dimensions, each having a different lengthscale parameter, is called the SE-ARD kernel:

$$\text{SE-ARD}(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^D \exp \sigma_d^2 \left( -\frac{1}{2} \frac{(x_d - x'_d)^2}{\ell_d^2} \right) = \sigma_f^2 \exp \left( -\frac{1}{2} \sum_{d=1}^D \frac{(x_d - x'_d)^2}{\ell_d^2} \right) \quad (2.4)$$

Figure 2.3 illustrates the SE-ARD kernel in two dimensions.

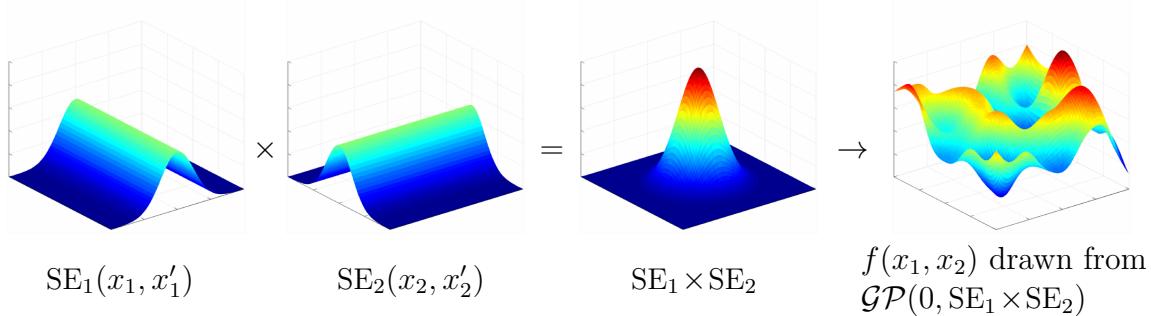


Figure 2.3 A product of two one-dimensional kernels gives rise to a prior on functions which depend on both dimensions.

ARD stands for Automatic Relevance Determination, so named because estimating the lengthscale parameters  $\ell_1, \ell_2, \dots, \ell_D$ , implicitly determines the “relevance” of each dimension. Input dimensions with relatively large lengthscales indicate relatively little variation along those dimensions in the function being modeled.

SE-ARD kernels are the default kernel in most applications of GPs. This may be partly because they have relatively few parameters to estimate, and those parameters are relatively interpretable. In addition, there is a theoretical reason to use them: they are *universal* kernels (Micchelli et al., 2006), capable of learning any continuous function given enough data, under some conditions.

However, this flexibility means that they can sometimes be relatively slow to learn, due to the *curse of dimensionality* (Bellman, 1956). In general, the more structure we

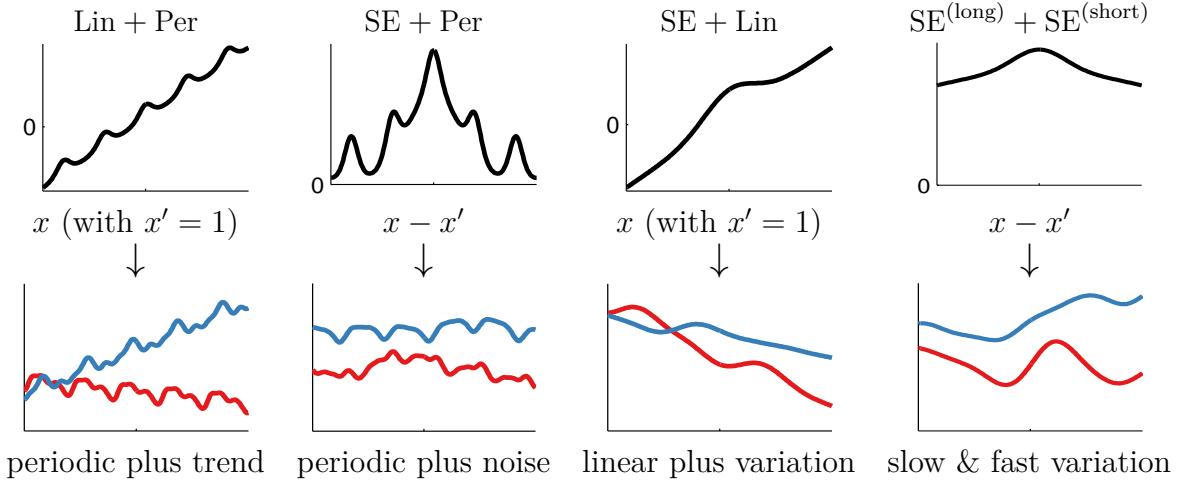


Figure 2.4 Examples of one-dimensional structures expressible by adding kernels. Rows have the same meaning as in figure 2.1.  $\text{SE}^{(\text{long})}$  denotes a SE kernel whose lengthscale is long relative to that of  $\text{SE}^{(\text{short})}$

account for, the less data we need - the *blessing of abstraction* (Goodman et al., 2011) counters the curse of dimensionality. Below, we will investigate ways to encode more structure into kernels.

## 2.4 Modeling sums of functions

An additive function is one which can be expressed as  $f(\mathbf{x}) = f_a(\mathbf{x}) + f_b(\mathbf{x})$ . Additivity is a very useful modeling assumption in a wide variety of contexts, especially if it allows us to make strong assumptions about the individual components which make up the sum. Restricting the flexibility of the component functions often aids in building interpretable models, and sometimes enables extrapolation in high dimensions.

Fortunately, it's easy to encode additivity into GP models. Suppose functions  $f_a, f_b$  are drawn independently from GP priors:

$$f_a \sim \mathcal{GP}(\mu_a, k_a) \quad (2.5)$$

$$f_b \sim \mathcal{GP}(\mu_b, k_b) \quad (2.6)$$

Then the sum of those functions is simply another GP:

$$f_a + f_b \sim \mathcal{GP}(\mu_a + \mu_b, k_a + k_b). \quad (2.7)$$

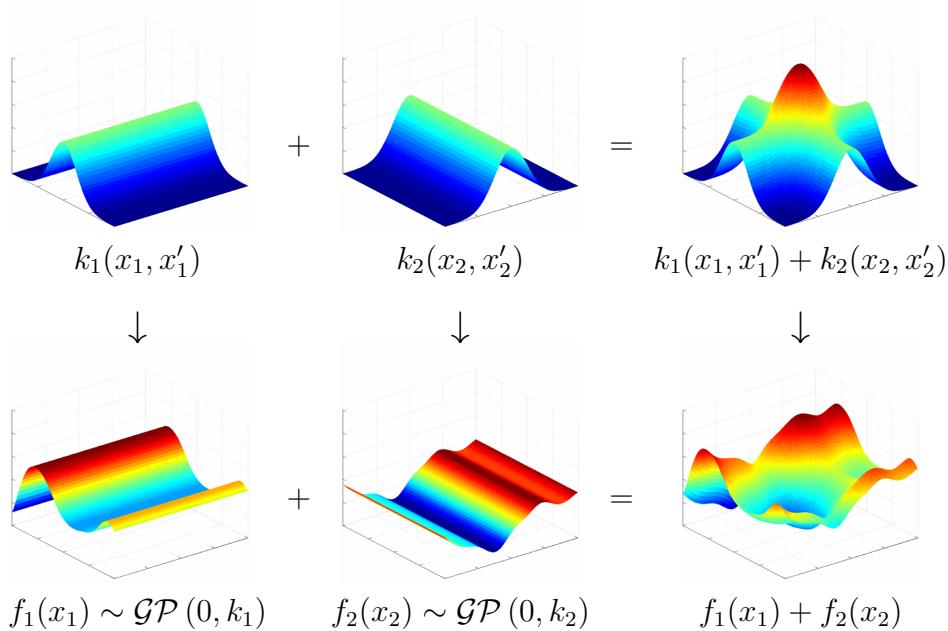


Figure 2.5 A sum of two orthogonal one-dimensional kernels. *Top row:* An additive kernel is any sum of kernels. *Bottom row:* A draw from an additive kernel corresponds to a sum of draws from independent GP priors, each having the corresponding kernel.

Kernels  $k_a$  and  $k_b$  can be kernels of different types, allowing us to model the data as a sum of independent functions, each possibly representing a different type of structure. We can also sum any number of components this way.

### 2.4.1 Additivity across multiple dimensions

When modeling functions of multiple dimensions, summing kernels can give rise to additive structure across different dimensions. To be more precise, if the kernels being added together are functions of only a subset of input dimensions, then the implied prior over functions decomposes in the same way. For example, if

$$f(x_1, x_2) \sim \mathcal{GP}(0, k_1(x_1, x'_1) + k_2(x_2, x'_2)) \quad (2.8)$$

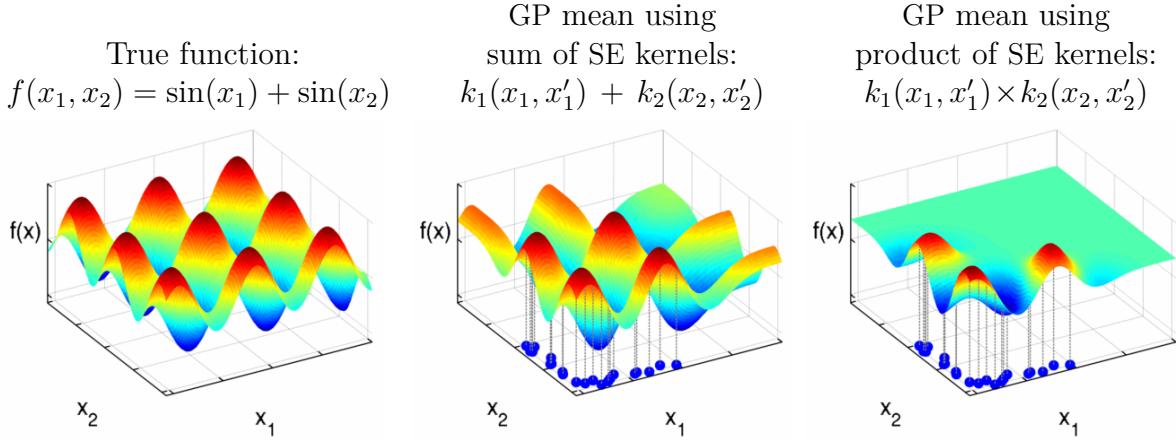


Figure 2.6 *Left*: A function with additive structure. *Center*: When the function being modeled has additive structure, a GP with an additive kernel can exploit this fact to extrapolate far from the training data. *Right*: The product kernel allows a different function value for every combination of inputs, and so is uncertain about function values away from the training data. This causes the predictions to revert to the mean.

Then this is equivalent to the model

$$f_1(x_1) \sim \mathcal{GP}(0, k_1(x_1, x'_1)) \quad (2.9)$$

$$f_2(x_2) \sim \mathcal{GP}(0, k_2(x_2, x'_2)) \quad (2.10)$$

$$f(x_1, x_2) = f_1(x_1) + f_2(x_2) \quad (2.11)$$

Figure 2.5 illustrates a decomposition of this form. Note that the product of two kernels does not have an analogous interpretation as the product of two functions.

## 2.4.2 Long-range extrapolation through additivity

Additive structure often allows us to make predictions far from the training data. Figure 2.6 compares the extrapolations made by additive versus product-kernel GP models, conditioned on data from a sum of two axis-aligned sine functions. The training data is confined to a small, L-shaped area. In this example, the additive model is able to correctly predict the height of the function at unseen combinations of inputs. The product-kernel model is more flexible, and so remains uncertain about the function away from the data.

These types of additive models have been well-explored in the statistics literature.

For example, generalized additive models (Hastie and Tibshirani, 1990) have seen wide adoption. In high dimensions, we can also consider sums of functions of more than one dimension. Chapter 6 considers this model class in more detail.

### 2.4.3 Example: an additive model of concrete strength

To illustrate how additive kernels give rise to interpretable models, we will build an additive model of the strength of concrete as a function of the amount of seven different ingredients, plus the age of the concrete (Yeh, 1998).

Our simple additive model looks like

$$\begin{aligned} f(\mathbf{x}) = & f_1(\text{cement}) + f_2(\text{slag}) + f_3(\text{fly ash}) + f_4(\text{water}) \\ & + f_5(\text{plasticizer}) + f_6(\text{coarse}) + f_7(\text{fine}) + f_8(\text{age}) + \text{noise} \end{aligned} \quad (2.12)$$

where noise  $\stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_n^2)$ . After learning the kernel parameters by maximizing the marginal likelihood of the data, we can visualize the predictive distribution of each component of the model.

Figure 2.7 shows the marginal posterior distribution of each of the eight components in equation (2.12). We can see that the parameters controlling the variance of two of the components, Coarse and Fine, were set to zero, meaning that the marginal likelihood preferred a parsimonious model which did not depend on these dimensions. This is an example of the automatic sparsity that arises by maximizing marginal likelihood in GP models, and another example of automatic relevance determination (ARD) (Neal, 1995).

The ability to learn kernel parameters in this way is much more difficult when using non-probabilistic methods such as Support Vector Machines (Cortes and Vapnik, 1995), for which cross-validation is often the best method to select kernel parameters.

### 2.4.4 Posterior variance of additive components

Here we derive the posterior variance and covariance of all of the additive components of a GP. These formulas allow us to make plots such as figure 2.7.

First, we will write down the joint prior over the sum of two functions drawn from GP priors. We distinguish between  $\mathbf{f}(\mathbf{X})$  (the function values at the training locations) and  $\mathbf{f}(\mathbf{X}^*)$  (the function values at some set of query locations).

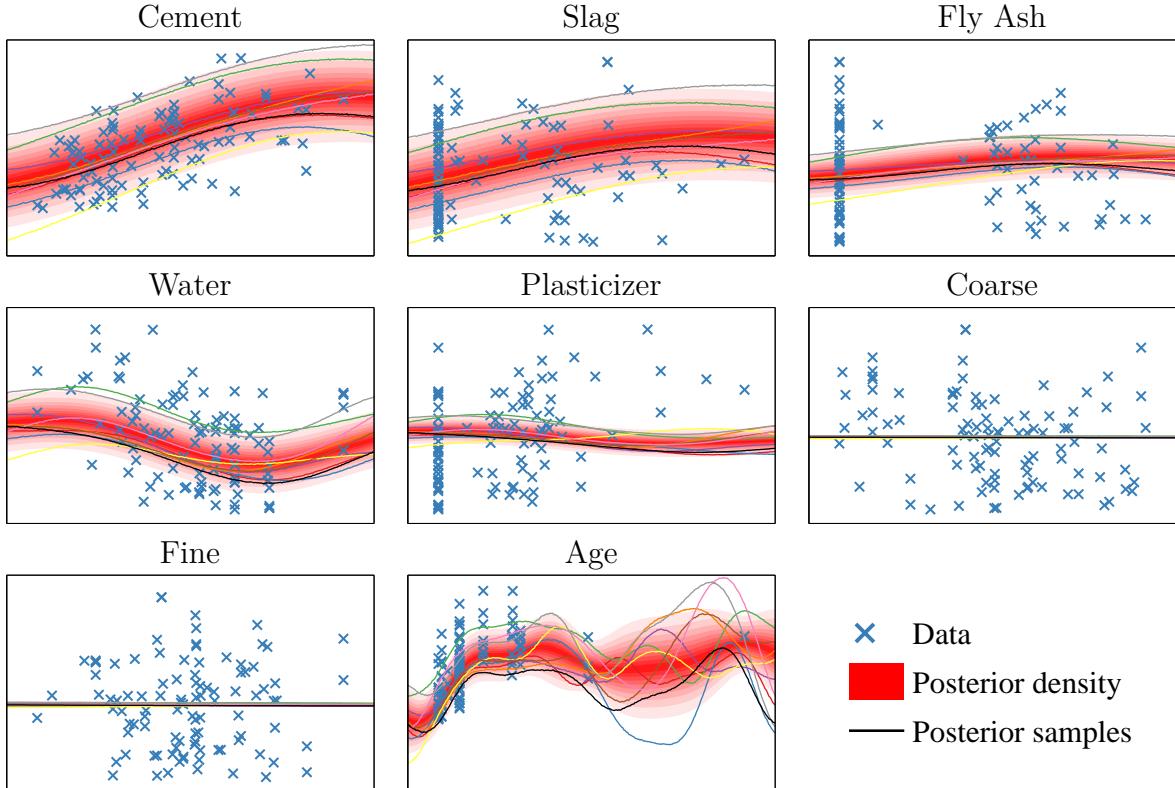


Figure 2.7 The predictive distribution of each component of a multidimensional additive model. Blue crosses indicate the original data projected on to each dimension, red indicates the marginal posterior density, and colored lines are samples from the marginal posterior of each component. The vertical axis is the same for all plots.

If  $f_1$  and  $f_2$  are *a priori* independent, and  $f_1 \sim \text{GP}(\mu_1, k_1)$  and  $f_2 \sim \text{GP}(\mu_2, k_2)$ , then

$$\begin{bmatrix} f_1(\mathbf{X}) \\ f_1(\mathbf{X}^*) \\ f_2(\mathbf{X}) \\ f_2(\mathbf{X}^*) \\ f_1(\mathbf{X}) + f_2(\mathbf{X}) \\ f_1(\mathbf{X}^*) + f_2(\mathbf{X}^*) \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_1^* \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2^* \\ \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^* \end{bmatrix}, \begin{bmatrix} \mathbf{K}_1 & \mathbf{K}_1^* & 0 & 0 & \mathbf{K}_1 & \mathbf{K}_1^* \\ \mathbf{K}_1^* & \mathbf{K}_1^{**} & 0 & 0 & \mathbf{K}_1^* & \mathbf{K}_1^{**} \\ 0 & 0 & \mathbf{K}_2 & \mathbf{K}_2^* & \mathbf{K}_2 & \mathbf{K}_2^* \\ 0 & 0 & \mathbf{K}_2^* & \mathbf{K}_2^{**} & \mathbf{K}_2^* & \mathbf{K}_2^{**} \\ \mathbf{K}_1 & \mathbf{K}_1^* & \mathbf{K}_2 & \mathbf{K}_2^* & \mathbf{K}_1 + \mathbf{K}_2 & \mathbf{K}_1^* + \mathbf{K}_2^* \\ \mathbf{K}_1^* & \mathbf{K}_1^{**} & \mathbf{K}_2^* & \mathbf{K}_2^{**} & \mathbf{K}_1^* + \mathbf{K}_2^* & \mathbf{K}_1^{**} + \mathbf{K}_2^{**} \end{bmatrix} \right) \quad (2.13)$$

where we represent the Gram matrices, evaluated at all pairs of vectors in bold capitals

as  $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ . So

$$\mathbf{K}_1 = k_1(\mathbf{X}, \mathbf{X}) \quad (2.14)$$

$$\mathbf{K}_1^* = k_1(\mathbf{X}^*, \mathbf{X}) \quad (2.15)$$

$$\mathbf{K}_1^{**} = k_1(\mathbf{X}^*, \mathbf{X}^*) \quad (2.16)$$

By the formula for Gaussian conditionals, (given by equation (A.2)), we get that the conditional distribution of a GP-distributed function conditioned on its sum with another GP-distributed function is given by

$$\begin{aligned} \mathbf{f}_1(\mathbf{X}^*) | \mathbf{f}_1(\mathbf{X}) + \mathbf{f}_2(\mathbf{X}) &\sim \mathcal{N}\left(\boldsymbol{\mu}_1^* + \mathbf{K}_1^{*\top} (\mathbf{K}_1 + \mathbf{K}_2)^{-1} [\mathbf{f}_1(\mathbf{X}) + \mathbf{f}_2(\mathbf{X}) - \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2] \right. \\ &\quad \left. \mathbf{K}_1^{**} - \mathbf{K}_1^{*\top} (\mathbf{K}_1 + \mathbf{K}_2)^{-1} \mathbf{K}_1^* \right) \end{aligned} \quad (2.17)$$

These formulas express the model's posterior uncertainty about the different components of the signal, integrating over the possible configurations of the other components. To extend these formulas to a sum of more than two functions, the term  $\mathbf{K}_1 + \mathbf{K}_2$  can simply be replaced by  $\sum_i \mathbf{K}_i$  everywhere.

### Posterior covariance of additive components

We can also compute the posterior covariance between any two components, conditioned on their sum: if this quantity is negative, it means that there is ambiguity about which of the two components explains the data at that location.

$$\text{cov} [\mathbf{f}_1(\mathbf{X}^*), \mathbf{f}_2(\mathbf{X}^*) | \mathbf{f}(\mathbf{X})] = -\mathbf{K}_1^{*\top} (\mathbf{K}_1 + \mathbf{K}_2)^{-1} \mathbf{K}_2^* \quad (2.18)$$

Figure 2.8 shows the posterior correlation between all non-zero components of the concrete model. Most of the correlation occurs within components, but there is also negative correlation between the “Cement” and “Slag” variables. This reflects an ambiguity in the model about which one of these functions is high and the other low.

## 2.5 Changepoints

A simple example of how combining kernels can give rise to more structured priors is given by changepoint kernels. Changepoints kernels can be defined through addition

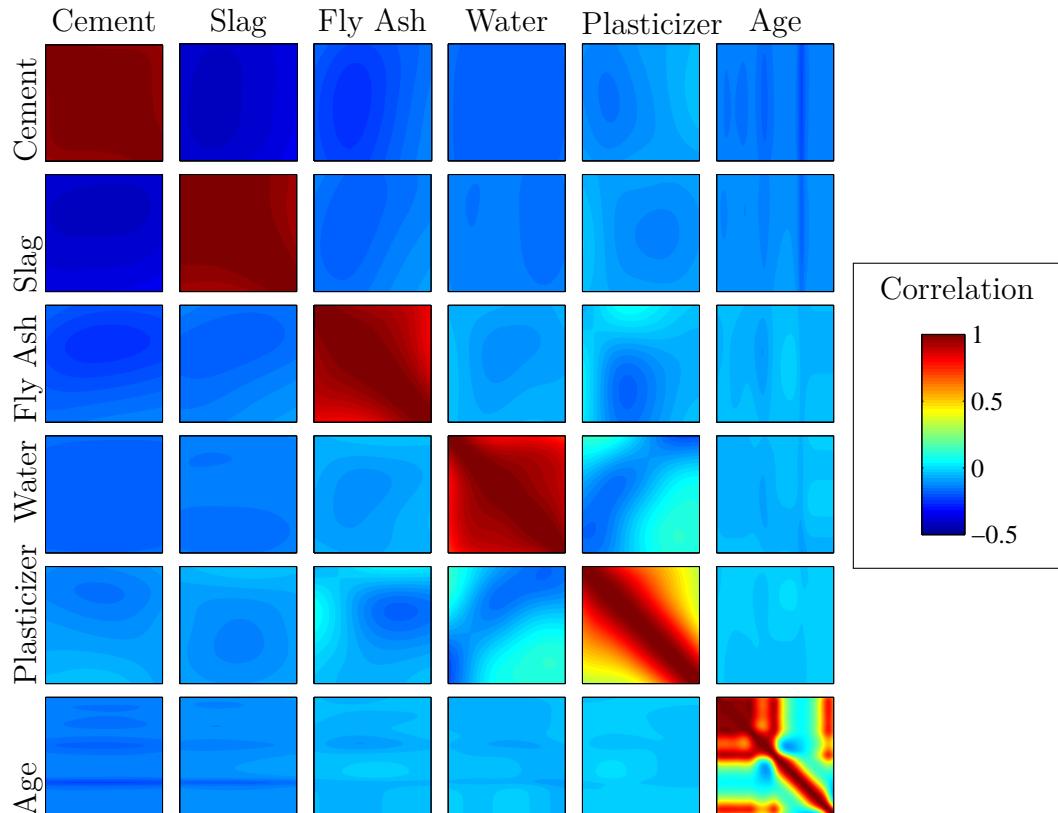


Figure 2.8 Posterior correlations between the components explaining the concrete dataset. Each plot shows the additive model's posterior correlations between two components, plotted over the domain of the data  $\pm 5\%$ . Red indicates high correlation, teal indicates no correlation, and blue indicates negative correlation. Plots on the diagonal show posterior correlations within each component. Dimensions 'Coarse' and 'Fine' are not shown, because their variance is zero.

and multiplication with sigmoidal functions such as  $\sigma(x) = 1/(1+\exp(-x))$ :

$$\text{CP}(k_1, k_2)(x, x') = \sigma(x)k_1(x, x')\sigma(x') + (1 - \sigma(x))k_2(x, x')(1 - \sigma(x')) \quad (2.19)$$

which can be written in shorthand as

$$\text{CP}(k_1, k_2) = k_1 \times \boldsymbol{\sigma} + k_2 \times \bar{\boldsymbol{\sigma}} \quad (2.20)$$

where  $\boldsymbol{\sigma} = \sigma(x)\sigma(x')$  and  $\bar{\boldsymbol{\sigma}} = (1 - \sigma(x))(1 - \sigma(x'))$ .

This compound kernel expresses a change from one kernel to another. The parameters of the sigmoid determine where, and how rapidly, this change occurs.

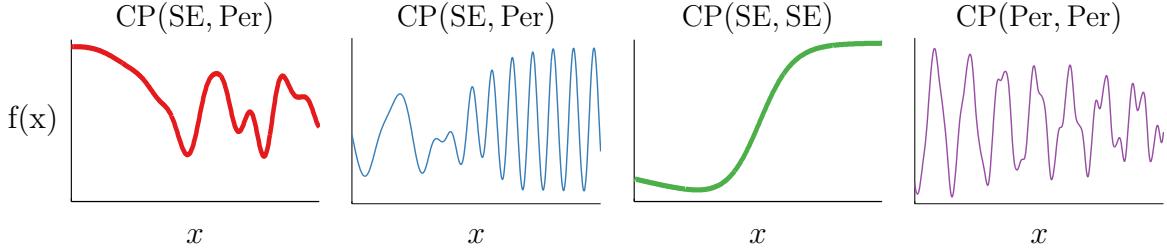


Figure 2.9 Draws from different priors on using changepoint kernels, constructed by adding and multiplying together base kernels with sigmoidal functions.

We can also express a function whose structure changes within some interval – a *change window* – by replacing  $\sigma(x)$  with a product of two sigmoids, one increasing and one decreasing.

### 2.5.1 Multiplication by a known function

More generally, we can model an unknown function that's been multiplied by some fixed, known function  $a(x)$ , by multiplying the kernel by  $a(\mathbf{x})a(\mathbf{x}')$ . Formally,

$$f(\mathbf{x}) = a(\mathbf{x})g(\mathbf{x}), \quad g \sim \mathcal{GP}(g \mid \mathbf{0}, k(\mathbf{x}, \mathbf{x}')) \iff f \sim \mathcal{GP}(f \mid \mathbf{0}, a(\mathbf{x})k(\mathbf{x}, \mathbf{x}')a(\mathbf{x}')) \quad (2.21)$$

## 2.6 Feature representation of kernels

By Mercer's theorem (Mercer, 1909), any positive-definite kernel can be represented as the inner product between a fixed set of features, evaluated at  $\mathbf{x}$  and at  $\mathbf{x}'$ :

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}') \quad (2.22)$$

As a simple example, the squared-exponential kernel (SE) on the real line has a representation in terms of infinitely many radial-basis functions of the form  $h_i(x) \propto \exp(-\frac{1}{2} \frac{(x-c_i)^2}{2\ell^2})$ . More generally, any stationary kernel on the real line can be represented by a set of sines and cosines - a Fourier representation (Bochner, 1959). In general, any particular feature representation of a kernel is not unique, and depends on which space  $\mathcal{X}$  is being considered (Minh et al., 2006).

In some cases,  $\mathcal{X}$  can even be the infinite-dimensional feature mapping of another kernel. Composing feature maps in this way leads to *deep kernels*, a topic explored in chapter 5.

### 2.6.1 Relation to linear regression

Surprisingly, GP regression is equivalent to Bayesian linear regression on the implicit features  $\mathbf{h}(\mathbf{x})$  which give rise to the kernel:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{h}(\mathbf{x}), \quad \mathbf{w} \sim \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{I}) \iff f \sim \mathcal{GP}(f | \mathbf{0}, \mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x})) \quad (2.23)$$

The link between Gaussian processes, linear regression, and neural networks is explored further in chapter 5.

### 2.6.2 Feature-space view of combining kernels

We can also view kernel addition and multiplication as a combination of the features of the original kernels. For example, if we have two kernels

$$k_a(\mathbf{x}, \mathbf{x}') = \mathbf{a}(\mathbf{x})^\top \mathbf{a}(\mathbf{x}') \quad (2.24)$$

$$k_b(\mathbf{x}, \mathbf{x}') = \mathbf{b}(\mathbf{x})^\top \mathbf{b}(\mathbf{x}') \quad (2.25)$$

then their addition has the form:

$$k_a(\mathbf{x}, \mathbf{x}') + k_b(\mathbf{x}, \mathbf{x}') = \mathbf{a}(\mathbf{x})^\top \mathbf{b}(\mathbf{x}') + \mathbf{a}(\mathbf{x})^\top \mathbf{b}(\mathbf{x}') = \begin{bmatrix} \mathbf{a}(\mathbf{x}) \\ \mathbf{b}(\mathbf{x}) \end{bmatrix}^\top \begin{bmatrix} \mathbf{a}(\mathbf{x}') \\ \mathbf{b}(\mathbf{x}') \end{bmatrix} \quad (2.26)$$

meaning that the features of  $k_a + k_b$  are the concatenation of the features of each kernel.

We can examine kernel multiplication in a similar way:

$$k_a(\mathbf{x}, \mathbf{x}') \times k_b(\mathbf{x}, \mathbf{x}') = [\mathbf{a}(\mathbf{x})^\top \mathbf{a}(\mathbf{x}')] \times [\mathbf{b}(\mathbf{x})^\top \mathbf{b}(\mathbf{x}')] \quad (2.27)$$

$$= \sum_i a_i(\mathbf{x}) a_i(\mathbf{x}') \times \sum_j b_j(\mathbf{x}) b_j(\mathbf{x}') \quad (2.28)$$

$$= \sum_{i,j} [a_i(\mathbf{x}) b_j(\mathbf{x})] [a_i(\mathbf{x}') b_j(\mathbf{x}')] \quad (2.29)$$

$$= \text{vec}(\mathbf{a}(\mathbf{x}) \otimes \mathbf{b}(\mathbf{x}'))^\top \text{vec}(\mathbf{a}(\mathbf{x}) \otimes \mathbf{b}(\mathbf{x}')) \quad (2.30)$$

In other words, the features of  $k_a \times k_b$  are just the Cartesian product (all possible combinations) of the original two sets of features. For example, the Cartesian product of the features of two one-dimensional SE kernels covers the plane with two-dimensional radial-basis functions of the form

$$h_{ij}(x_1, x_2) \propto \exp\left(-\frac{1}{2} \frac{(x_1 - c_i)^2}{2\ell_1^2}\right) \exp\left(-\frac{1}{2} \frac{(x_2 - c_j)^2}{2\ell_2^2}\right) \quad (2.31)$$

## 2.7 Expressing symmetries and invariants

When modeling functions, encoding known symmetries can improve predictive accuracy. This section looks at different ways to encode symmetries into a prior on functions. Many types of symmetry can be enforced through operations on the kernel.

We will demonstrate the properties of the resulting models by sampling functions from their priors. By using these functions to define warpings from  $\mathbb{R}^2 \rightarrow \mathbb{R}^3$ , we will show how to build a nonparametric prior on an open-ended family of topological manifolds, such as cylinders, toruses, and Möbius strips.

### 2.7.1 Three recipes for invariant priors

Consider the scenario where we have a finite set of transformations of the input space  $\{g_1, g_2, \dots\}$  to which we wish our function to remain invariant:

$$f(\mathbf{x}) = f(g(\mathbf{x})) \quad \forall \mathbf{x} \in \mathcal{X}, \quad \forall g \in G \quad (2.32)$$

As an example, imagine we want to build a model of functions invariant to swapping their inputs:  $f(x_1, x_2) = f(x_2, x_1)$ . Being invariant to a set of operations is equivalent to being invariant to all compositions of those operations, the set of which form a group. (Armstrong et al., 1988, chapter 21). In our example, the elements of the group  $G_{\text{swap}}$  containing the operations the functions are invariant to has two elements:

$$g_1([x_1, x_2]) = [x_2, x_1] \quad (\text{swap}) \quad (2.33)$$

$$g_2([x_1, x_2]) = [x_1, x_2] \quad (\text{identity}) \quad (2.34)$$

How can we construct a prior on functions which respect these symmetries?

Ginsbourger et al. (2012) and Ginsbourger et al. (2013) showed that the only way to construct a GP prior on functions which respect a set of invariants is to construct a kernel which respects the same invariants with respect to each of its two inputs:

$$k(\mathbf{x}, \mathbf{x}') = k(g(\mathbf{x}), g(\mathbf{x}')), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \quad \forall g \in G \quad (2.35)$$

Formally, given a finite group  $G$  whose elements are operations to which we wish our function to remain invariant, and  $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ , then every  $f$  is invariant under  $G$  (up to a modification) if and only if  $k(\cdot, \cdot)$  is argument-wise invariant under  $G$ .

It might not always be clear how to construct a kernel respecting such argument-wise invariances. Fortunately, there are a few simple ways to do this for any finite group:

1. **Sum over the orbit.** Ginsbourger et al. (2012) and Kondor (2008) suggest enforcing invariants through a double sum over the orbits of  $\mathbf{x}$  and  $\mathbf{x}'$  with respect to  $G$ :

$$k_{\text{sum}}(\mathbf{x}, \mathbf{x}') = \sum_{g, \in G} \sum_{g' \in G} k(g(\mathbf{x}), g'(\mathbf{x}')) \quad (2.36)$$

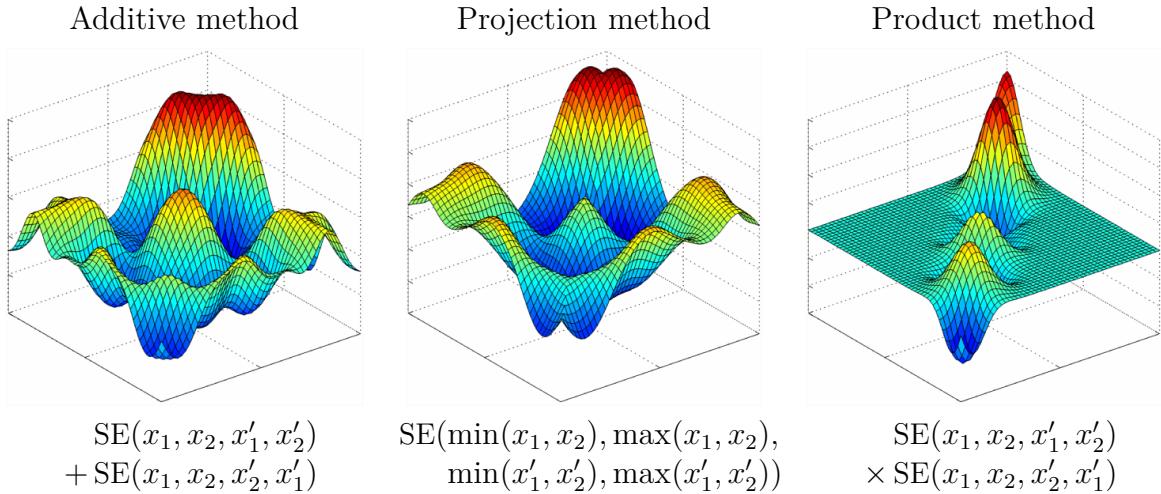


Figure 2.10 Three methods of introducing symmetry, illustrated through draws from the corresponding priors. All three methods introduce a different type of nonstationarity.

For the group  $G_{\text{swap}}$ , this operation results in the kernel:

$$k_{\text{switch}}(\mathbf{x}, \mathbf{x}') = \sum_{g \in G_{\text{swap}}} \sum_{g' \in G_{\text{swap}}} k(g(\mathbf{x}), g'(\mathbf{x}')) \quad (2.37)$$

$$\begin{aligned} &= k(x_1, x_2, x'_1, x'_2) + k(x_1, x_2, x'_2, x'_1) \\ &\quad + k(x_2, x_1, x'_1, x'_2) + k(x_2, x_1, x'_2, x'_1) \end{aligned} \quad (2.38)$$

For stationary kernels, some pairs of elements in this sum will be identical, and can be ignored. Figure 2.10(left) shows a draw from a GP prior with an SE kernel symmetrized in this way. This construction has the property that the marginal variance is doubled near  $x = y$ , which may or may not be desirable.

2. **Project onto a fundamental domain.** Ginsbourger et al. (2013) also explore the possibility of projecting each datapoint into a fundamental domain of the group, using a mapping  $A_G$ :

$$k_{\text{proj}}(\mathbf{x}, \mathbf{x}') = k(A_G(\mathbf{x}), A_G(\mathbf{x}')) \quad (2.39)$$

For the group  $G_{\text{swap}}$ , a fundamental domain is  $\{x_1, x_2 : x_1 < x_2\}$ , which can be mapped to using  $A_{G_{\text{swap}}}(x_1, x_2) = [\min(x_1, x_2), \max(x_1, x_2)]$ . Constructing a kernel using this method introduces a non-differentiable “seam” along  $x_1 = x_2$ , as shown in figure 2.10(center).

**3. Multiply over the orbit.** Ryan Adams (personal communication) suggested a construction enforcing invariants through products over the orbits:

$$k_{\text{sum}}(\mathbf{x}, \mathbf{x}') = \prod_{g \in G} \prod_{g' \in G} k(g(\mathbf{x}), g'(\mathbf{x}')) \quad (2.40)$$

This method can sometimes produce GP priors with zero variance in some regions of the space, as in figure 2.10(right).

There are many possible ways to achieve a given symmetry, but we must be careful to do so without compromising other qualities of the model we are constructing. For example, simply setting  $k(\mathbf{x}, \mathbf{x}') = 0$  gives rise to a GP prior which obeys *all possible* symmetries, but this is presumably not a model we wish to use.

### 2.7.2 Example: periodicity

Periodicity in a one-dimensional function corresponds to the invariance

$$f(x) = f(x + \tau) \quad (2.41)$$

where  $\tau$  is the period.

The most popular method for building a periodic kernel is due to MacKay (1998), who used the projection method in combination with an SE kernel. A fundamental domain of the symmetry group is a circle, so the kernel

$$\text{Per}(x, x') = \text{SE}\left(\left[\sin(x), \cos(x)\right], \left[\sin(x'), \cos(x')\right]\right) \quad (2.42)$$

achieves the invariance in equation (2.41). Simple algebra reduces this kernel to the form shown in figure 2.1.

We could also build a periodic kernel with period  $\tau$  by the mapping  $A(x) = \text{mod}(x, \tau)$ . However, samples from this prior would be discontinuous at every integer multiple of  $\tau$ .

### 2.7.3 Example: symmetry about zero

Another simple example of an easily-enforceable symmetry is symmetry about zero:

$$f(x) = f(-x) \quad (2.43)$$

using the sum over orbits method, by the transform

$$k_{\text{reflect}}(x, x') = k(x, x') + k(x, -x') + k(-x, x') + k(-x, -x') \quad (2.44)$$

### 2.7.4 Example: translation invariance in images

Most models of images are invariant to spatial translations (LeCun and Bengio, 1995). Similarly, most models of sounds are also invariant to translation through time.

This sort of translation invariance is completely distinct from the stationarity of kernels such as SE or Per. A stationary kernel implies that the prior is invariant to translations of the entire training and test set. In contrast, we use translation invariance to refer to the situation where the signal has been discretized, and each pixel (or the audio equivalent) corresponds to a different input dimension. We are interested in creating priors on functions that are invariant to swapping pixels in a manner that corresponds to shifting in some direction:

$$f\left(\begin{array}{|c|c|c|}\hline & \blacksquare & \\ \hline \blacksquare & & \\ \hline \end{array}\right) = f\left(\begin{array}{|c|c|c|}\hline & & \blacksquare \\ \hline & \blacksquare & \\ \hline \end{array}\right) \quad (2.45)$$

For example, in a one-dimensional image or audio signal, translation of an input vector by  $i$  pixels can be defined as

$$\text{shift}(\mathbf{x}, i) = [x_{\text{mod}(i+1, D)}, x_{\text{mod}(i+2, D)}, \dots, x_{\text{mod}(i+D, D)}]^T \quad (2.46)$$

As above, translation invariance in one dimension can be achieved by the transformation

$$k_{\text{invariant}}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^D \sum_{j=1}^D k(\text{shift}(\mathbf{x}, i), \text{shift}(\mathbf{x}', j)), \quad (2.47)$$

simply defining the covariance between two signals to be the sum of all covariances between all translations of those two signals.

The extension to two dimensions,  $\text{shift}(\mathbf{x}, i, j)$ , is straightforward, but notationally cumbersome. Kondor (2008) built a more elaborate kernel between images, approximately invariant to both translation and rotation by using the projection method.

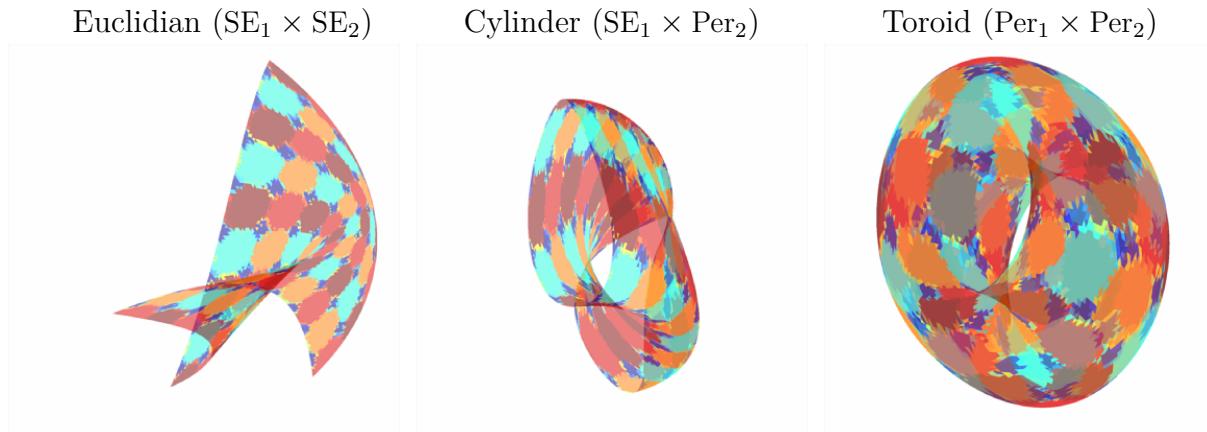


Figure 2.11 Generating 2D manifolds with different topologies. By enforcing that the functions mapping from  $\mathbb{R}^2$  to  $\mathbb{R}^3$  obey the appropriate symmetries, the surfaces created have the corresponding topologies, ignoring self-intersections.

## 2.8 Generating topological manifolds

In this section we give a geometric illustration of the symmetries encoded by different compositions of kernels. The work presented in this section is based on a collaboration with David Reshef, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. The derivation of the Möbius kernel was an original contribution by myself.

Priors on functions exhibiting symmetries can be used to create a prior on topological manifolds, by warping a simply-connected surface into a higher-dimensional space using a mapping expressing symmetries. To build a prior on 2-dimensional manifolds embedded in 3-dimensional space, we simply need a prior on mappings from  $\mathbb{R}^2$  to  $\mathbb{R}^3$ , which we can construct using three independent functions  $[f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})]$ , each mapping from  $\mathbb{R}^2$  to  $\mathbb{R}$ . GP priors on the mapping functions implicitly give rise to a prior on warped surfaces. Symmetries in  $[f_1, f_2, f_3]$  will connect different parts of the manifolds, giving rise to non-trivial topologies on the sampled surfaces. Figure 2.11 shows 2D meshes warped into 3D by functions drawn from GP priors with different kernels, giving rise to a variety of different topologies. Higher-dimensional analogues of these shapes can be constructed by increasing the latent dimension and including corresponding terms in the kernel. For example, an  $N$ -dimensional space with kernel  $\text{Per}_1 \times \text{Per}_2 \times \dots \times \text{Per}_N$  will give rise to a prior on  $N$ -dimensional toruses.

This construction is similar in spirit to the GP latent variable model (GP-LVM) of Lawrence (2005), which learns a latent embedding of the data into a low-dimensional space, using a GP prior on the mapping from the latent space to the observed space.

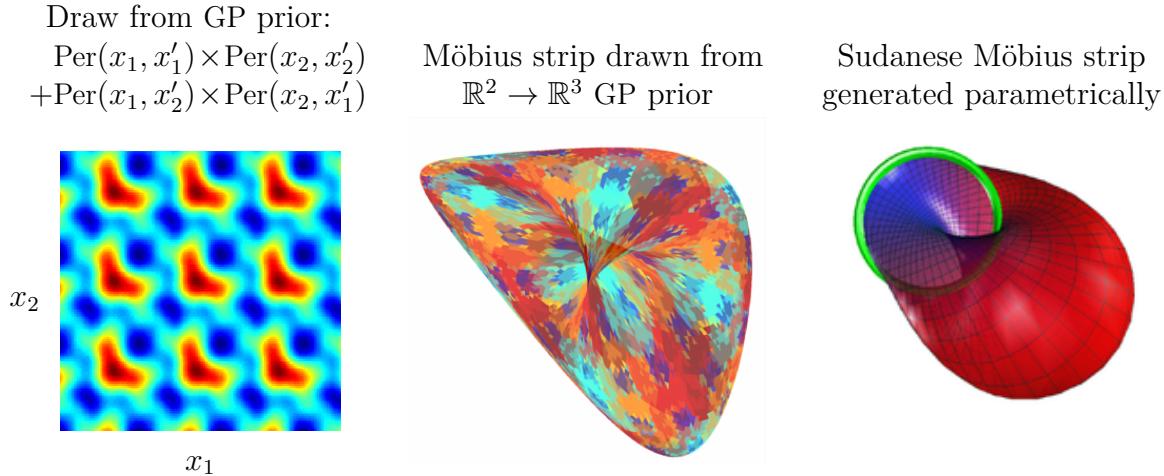


Figure 2.12 Generating Möbius strips. *Left:* A function drawn from a GP prior obeying the same symmetries as a Möbius strip. *Center:* By enforcing that the functions mapping from  $\mathbb{R}^2$  to  $\mathbb{R}^3$  obey the appropriate symmetries, surfaces sampled from the prior have topology corresponding to a Möbius strip. Surfaces generated this way do not have the familiar shape of a flat surface connected to itself with a half-twist. Instead, they tend to look like *Sudanese Möbius strips* (Lerner and Asimov, 1984), whose edge has a circular shape. *Right:* A Sudanese projection of a Möbius strip. Image adapted from [Wikimedia Commons](#) (2005).

### 2.8.1 Möbius strips

A prior on functions on Möbius strips can be constructed by enforcing the symmetries:

$$\begin{aligned} f(x_1, x_2) &= f(x_1 + \tau, x_2) && (\text{periodic in } x_1) \\ f(x_1, x_2) &= f(x_1, x_2 + \tau) && (\text{periodic in } x_2) \\ f(x_1, x_2) &= f(x_2, x_1) && (\text{symmetric about } x_1 = x_2) \end{aligned}$$

Moving along the diagonal  $x_1 = x_2$  of a function drawn from the corresponding GP prior is equivalent to moving along the edge of a notional Möbius strip which has had the function mapped on to its surface. Figure 2.12(a) shows an example of a function drawn from such a prior. Figure 2.12(b) shows an example of a 2D mesh mapped to 3D by functions drawn from such a prior. This surface doesn't resemble the typical representation of a Möbius strip, but instead resembles an embedding known as the Sudanese Möbius strip (Lerner and Asimov, 1984), shown in figure 2.12(c).

## 2.9 Kernels on categorical variables

Categorical variables are variables which can take values only from a discrete, unordered set, such as  $\{\text{blue}, \text{green}, \text{red}\}$ . A flexible way to construct a kernel over categorical variables is to represent that variable by a set of binary variables, using a one-of-k encoding. For example, if  $\mathbf{x}$  can take one of four values,  $x \in \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ , then a one-of-k encoding of  $x$  will correspond to four binary inputs, and  $\text{one-of-k}(\mathbf{C}) = [0, 0, 1, 0]$ . Given a one-of-k encoding, we can place any multidimensional kernel on that space, such as the SE-ARD:

$$k_{\text{categorical}}(x, x') = \text{SE-ARD}(\text{one-of-k}(x), \text{one-of-k}(x')) \quad (2.48)$$

Short lengthscales on any particular dimension in the SE-ARD kernel indicate that the function value corresponding to that category is uncorrelated with the others.

A more flexible parameterization suggested by Kevin Swersky (personal communication) allows complete flexibility about which pairs of categories are similar to one another, replacing the SE-ARD kernel with a fully-parameterized kernel, SE-full:

$$\text{SE-full}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{L} \mathbf{x}'\right) \quad (2.49)$$

where  $L$  is a symmetric matrix, individually parameterizing the covariance between each pair of function values of categories.

## 2.10 Building a kernel in practice

One difficulty in building GP models is choosing, or integrating over, the kernel parameters. If the kernel we construct has relatively few parameters, these parameters can be estimated by maximum marginal likelihood, using gradient-based optimizers. The kernel parameters estimated in the examples above were optimized using the GPML toolbox, available at [gaussianprocess.org/gpml/code](http://gaussianprocess.org/gpml/code).

The next chapter will show how to perform such a search not just over the kernel parameters, but also over the open-ended space of kernel expressions.

### Source code

Source code to produce all figures is available at [github.com/duvenaud/phd-thesis](https://github.com/duvenaud/phd-thesis).

# Chapter 3

## Automatic Model Construction

“It would be very nice to have a formal apparatus that gives us some ‘optimal’ way of recognizing unusual phenomena and inventing new classes of hypotheses that are most likely to contain the true one; but this remains an art for the creative human mind.”

– E. T. Jaynes (1985)

In chapter 2, we saw that the choice of kernel determines the type of structure that can be learned by a GP model, and that a wide variety of models could be constructed through simply adding and multiplying a few base kernels together. However, we did not answer the difficult question of which kernel to use for a given problem. Even for experts, choosing the kernel in GP regression remains something of a black art.

The contribution of this chapter is to show how to automate the process of building kernels for GP models. To do so, we define an open-ended space of kernels, by adding and multiplying together kernels from a fixed set. We then search over this space to find a kernel which captures as much structure in the data as possible.

Searching over such a large, structured model class has two main benefits. First, this procedure has very good predictive accuracy, since it tries out a large number of different regression models. Second, this procedure can discover interpretable structure in datasets. Because GP posteriors can be decomposed (as in section 2.4.3), we can also examine the resulting structures visually. In chapter 4, we also show how to automatically generate English-language descriptions of the resulting models.

This chapter is based on work done in collaboration with James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. It was published in [Duvenaud et al. \(2013\)](#) and [Lloyd et al. \(2014\)](#). Myself, James Lloyd and Roger Grosse jointly developed the idea of searching through a grammar-based language of GP models, inspired

by Grosse et al. (2012), and wrote the first versions of the code together. James Lloyd ran most of the experiments, while I constructed most of the figures.

## 3.1 Ingredients of an automatic statistician

Gelman (2013) asks, “How can an artificial intelligence do statistics? ... It needs not just an inference engine, but also a way to construct new models and a way to check models. Currently, those steps are performed by humans, but the AI would have to do it itself.”. This section will discuss the different parts we think are required to build an artificial intelligence that can do statistics.

- 1. An open-ended language of models.** Many learning algorithms consider all models in a class of fixed size. For example, graphical model learning algorithms (Eaton and Murphy, 2007; Friedman and Koller, 2003) search over different connectivity graphs for a given set of nodes. While such methods can be powerful, human statisticians are capable of deriving novel model classes when required. An automatic search through an open-ended class of models can achieve some of this flexibility, growing the complexity of the model as needed, possibly combining existing structures in novel ways.
- 2. A search through model space.** An open-ended space of models cannot be searched exhaustively. Just as human researchers iteratively refine their models, search procedures can propose new search directions based on the results of previous model fits. Because any search in an open-ended space must start with relatively simple models before moving on to more complex ones, any search strategy is likely to resemble an iterative model-building procedure.
- 3. A model comparison procedure.** A search strategy requires an objective to optimize. In this work, we use approximate marginal likelihood to compare models, penalizing complexity using the Bayesian Information Criterion as a heuristic. More generally, an automatic statistician should be able to question the models it has constructed. Gelman and Shalizi (2012) review the literature on model checking.
- 4. A model description procedure.** Part of the value of statistical models comes from helping humans to understand a dataset or a phenomenon. Furthermore, a clear description of the statistical structure found in a dataset helps a user to notice when the

dataset has errors, the wrong question was asked, the model-building procedure failed to capture known structure, a relevant piece of data or constraint is missing, or when a novel statistical structure has been found.

In this chapter, we introduce a system containing all the above ingredients. We call this system the automatic Bayesian covariance discovery (ABCD) system. The next four sections of this chapter describe the mechanisms we use to produce these four ingredients, for this particular example of an artificial intelligence which does statistics.

## 3.2 A language of regression models

As shown in chapter 2, we can construct a wide variety of kernel structures compositionally by adding and multiplying a small number of base kernels. We can therefore define a language of GP regression models simply by specifying a language of kernels.

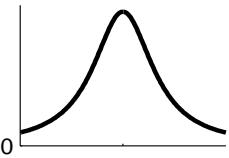
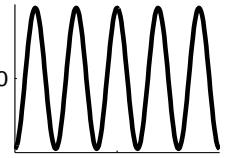
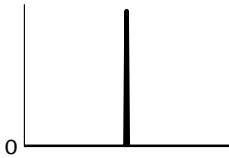
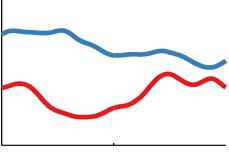
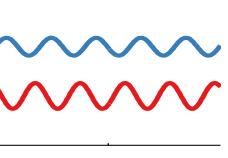
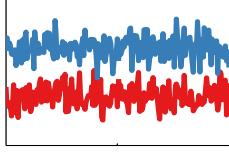
Kernel name:	Rational quadratic (RQ)	Cosine (cos)	White noise (Lin)
$k(x, x') =$	$\left(1 + \frac{(x-x')^2}{2\alpha\ell^2}\right)^{-\alpha}$	$\cos\left(2\pi\frac{(x-x')}{p}\right)$	$\delta(x - x')$
Plot of kernel:			
Samples from GP prior:			
Type of structure:	multiscale variation	sine waves	uncorrelated noise

Figure 3.1 New base kernels introduced in this chapter, and the types of structure they encode. More interesting kernels can be constructed by adding and multiplying base kernels together.

To specify an open-ended language of structured kernels, we will consider the set of

all kernels that can be built by adding and multiplying these base kernels together:

$$k_1 + k_2 = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$
 (3.1)

$$k_1 \times k_2 = k_1(\mathbf{x}, \mathbf{x}') \times k_2(\mathbf{x}, \mathbf{x}')$$
 (3.2)

This language of models is made out of a set of base kernels which capture different properties of functions, and a set of composition rules which combine kernels to yield other valid kernels. In this chapter, we will use such base kernels as white noise (WN), constant (C), linear (Lin), squared-exponential (SE), rational-quadratic (RQ), sigmoidal ( $\sigma$ ) and periodic (Per). We use a generalized form of Per due to James Lloyd (personal communication) which has  $\cos(x - x')$  as a special case. Figure 3.1 shows the new kernels introduced in this chapter. For precise definitions of all kernels, see appendix B.

The space of kernels produced by adding and multiplying the above set of kernels contains many existing regression models. Table 3.1 lists some of these, which are discussed in more detail in section 3.7.

Regression model	Kernel	Related work
Linear regression	C + Lin + WN	
Polynomial regression	C + $\prod$ Lin + WN	
Semi-parametric	Lin + SE + WN	Ruppert et al. (2003)
Multiple kernel learning	$\sum$ SE + WN	Gönen and Alpaydin (2011)
Trend, cyclical, irregular	$\sum$ SE + $\sum$ Per + WN	Lind et al. (2006)
Fourier decomposition	C + $\sum$ cos + WN	
Sparse spectrum GPs	$\sum$ cos + WN	Lázaro-Gredilla et al. (2010)
Spectral mixture	$\sum$ SE $\times$ cos + WN	Wilson and Adams (2013)
Changepoints	e.g. CP(SE, SE) + WN	Garnett et al. (2010)
Time-changing variance	e.g. SE + Lin $\times$ WN	
Interpretable + flexible	$\sum_d$ SE <sub>d</sub> + $\prod_d$ SE <sub>d</sub>	Plate (1999)
Additive GPs	$\sum_{\mathbf{r} \in \{0,1\}^D} \prod_{d \in \mathbf{r}}$ SE <sub>d</sub>	chapter 6

Table 3.1 Common regression models expressible by sums and products of base kernels.  $\cos(\cdot, \cdot)$  is a special case of our reparametrised Per( $\cdot, \cdot$ ).

### 3.3 A model search procedure

We explore the space of regression models using a simple greedy search. At each stage, we choose the highest scoring kernel, and propose modifying it by applying an operation

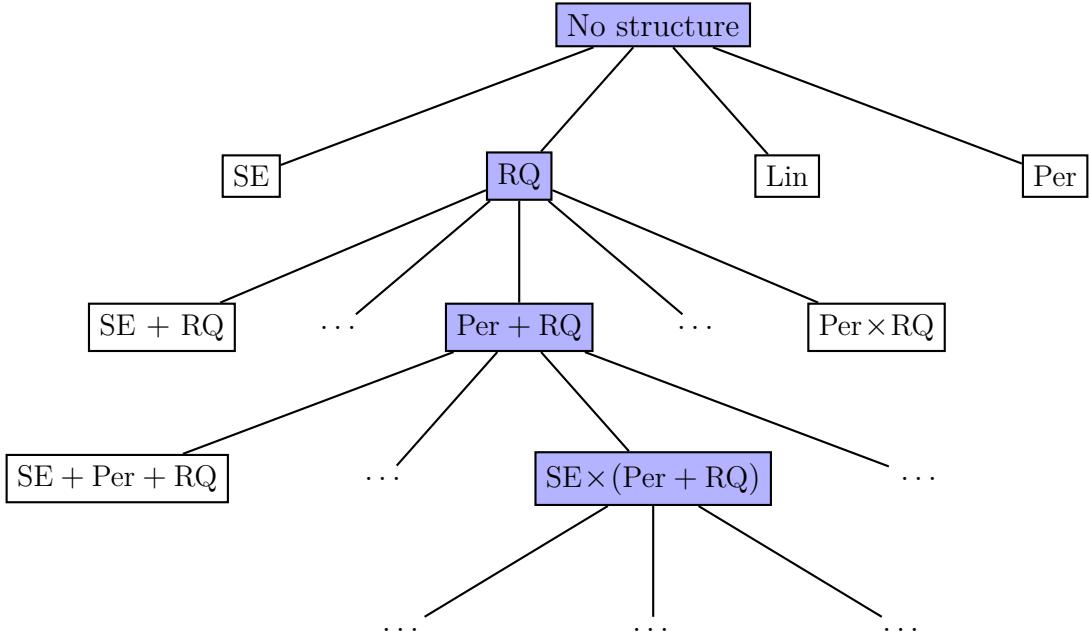


Figure 3.2 An example of a search tree over kernel expressions. Figure 3.3 shows the corresponding model increasing in sophistication as the kernel expression grows.

to one of its parts, combining or replacing that part with another base kernel. The basic operations we can perform on any part  $k$  of a kernel are:

$$\begin{aligned} \text{Replacement: } k &\rightarrow k' \\ \text{Addition: } k &\rightarrow k + k' \\ \text{Multiplication: } k &\rightarrow k \times k' \end{aligned}$$

where  $k'$  is a new base kernel. These operators can generate all possible algebraic expressions involving addition and multiplication of base kernels. To see this, observe that if we restricted the addition and multiplication rules to only apply to base kernels, we would obtain a context-free grammar which generates the set of algebraic expressions.

Figure 3.2 shows an example search tree followed by our algorithm. Figure 3.3 shows how the resulting model changes as the search is followed. In practice, we also include extra operators which propose commonly-occurring structures, such as changepoints. A complete list is contained in appendix C.

Our search operators are motivated by strategies that human researchers often use to construct kernels. In particular,

- One can look for structure, such as periodicity, in the residuals of a model, and then extend the model to capture that structure. This corresponds to adding a new kernel to the existing structure.

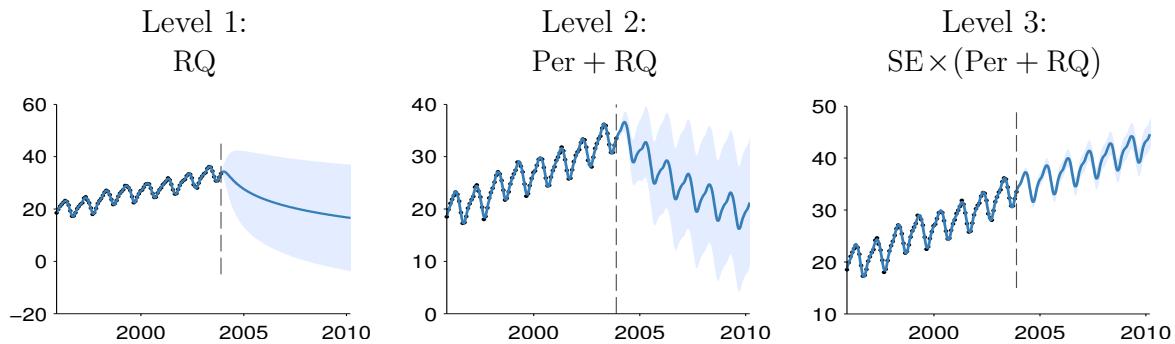


Figure 3.3 Posterior mean and variance for different depths of kernel search on the Mauna Loa dataset. The dashed line marks the end of the dataset. *Left:* the function is only modeled as a locally smooth function, and the extrapolation is poor. *Middle:* a periodic component is added, and the extrapolation improves. *Right:* at depth 3, the kernel can capture most of the relevant structure, and is able to extrapolate reasonably.

- One can start with structure, such as linearity, which is assumed to hold globally, but find that it only holds locally. This corresponds to multiplying a kernel structure by a local kernel, such as SE.
- One can add features incrementally, analogous to algorithms like boosting, back-fitting, or forward selection. This corresponds to adding or multiplying with kernels on dimensions not yet included in the model.

## Hyperparameter initialization

Unfortunately, optimizing the marginal likelihood over parameters is not a convex optimization problem, and the space can have many local optima. For example, in data with periodic structure, integer multiples of the true period (harmonics) are often local optima. We take advantage of our search procedure to provide reasonable initializations: all of the parameters which were part of the previous kernel are initialized to their previous values, while randomly initializing any newly introduced parameters. In the newly proposed kernel, all parameters are then optimized using conjugate gradients. This procedure is not guaranteed to find the global optimum, but it implements the commonly used heuristic of iteratively modeling residuals.

### 3.4 A model comparison procedure

Choosing a kernel requires a method for comparing models. We choose marginal likelihood as our criterion, since it balances the fit and complexity of a model (Rasmussen and Ghahramani, 2001). Conditioned on kernel parameters, the marginal likelihood of a GP can be computed analytically. Given a parametric form of a kernel, we can also choose its parameters using marginal likelihood. However, choosing kernel parameters by maximum likelihood (as opposed to integrating them out) raises the possibility of overfitting. In addition, if we compare two classes of kernels by the maximum likelihood found by optimizing over the kernel parameters, then all else being equal, the kernel class having more free parameters will be chosen. This introduces a bias for more complex models.

We could avoid overfitting by integrating the marginal likelihood over all free parameters, but this integral is difficult to do in general. Instead, we approximate this integral using the Bayesian information criterion (BIC) (Schwarz, 1978):

$$\text{BIC}(M) = \log p(D | M) - \frac{1}{2}|M| \log N \quad (3.3)$$

where  $p(D|M)$  is the marginal likelihood of the data (given by equation (1.3)),  $|M|$  is the number of kernel parameters, and  $N$  is the number of data points. BIC simply penalizes the marginal likelihood in proportion to how many parameters the model has. Because BIC is a function of the number of parameters in a model, we did not count kernel parameters known to not affect the model. For example, when two kernels are multiplied, one of their output variance parameters becomes redundant, and can be ignored.

Other more sophisticated approximations are possible, such as Laplace's approximation. We chose to try BIC first because of its simplicity, and it performed reasonably in our experiments.

### 3.5 A model description procedure

As discussed in Chapter 2, a GP whose kernel is a sum of kernels can be viewed as a sum of functions drawn from different GPs. We can always express any kernel structure

as a sum of products of kernels, by distributing all products of sums. For example,

$$\text{SE} \times (\text{RQ} + \text{Lin}) = \text{SE} \times \text{RQ} + \text{SE} \times \text{Lin} \quad (3.4)$$

This decomposition into additive components provides a method of visualizing the learned model, breaking down the different types of structure discovered in the data.

For products of kernels which all apply to the same dimension, we can use the formulas in section 2.4.4 to visualize the marginal posterior distribution of each component. The following section shows examples of such decomposition plots. In chapter 4, we extend this model visualization method to include automatically generated English text explaining the meaning of each type of structure discovered.

## 3.6 Structure discovery in time series

To investigate our method’s ability to discover structure, we ran the kernel search on several time-series. In the following examples, the search was run to depth 10, using SE, RQ, Lin, Per and WN as base kernels.

### 3.6.1 Mauna Loa atmospheric CO<sub>2</sub>

Using our method, we analyzed records of carbon dioxide levels recorded at the Mauna Loa observatory. Since this dataset was analyzed in detail by [Rasmussen and Williams \(2006\)](#), we can compare the kernel chosen by our method to a kernel constructed by human experts.

Figure 3.3 shows the posterior mean and variance on this dataset as the search depth increases. While the data can be smoothly interpolated by a model with only a single base kernel, the extrapolations improve dramatically as the increased search depth allows more structure to be included.

Figure 3.4 shows the final model chosen by our method, together with its decomposition into additive components. The final model exhibits both plausible extrapolation and interpretable components: a long-term trend, annual periodicity, and medium-term deviations. These are the same components chosen in the kernel hand-constructed by [Rasmussen and Williams \(2006, Chapter 5\)](#).

We also plot the residuals modeled by a white noise (WN) component, showing that there is little obvious structure left in the data. More generally, some components capture

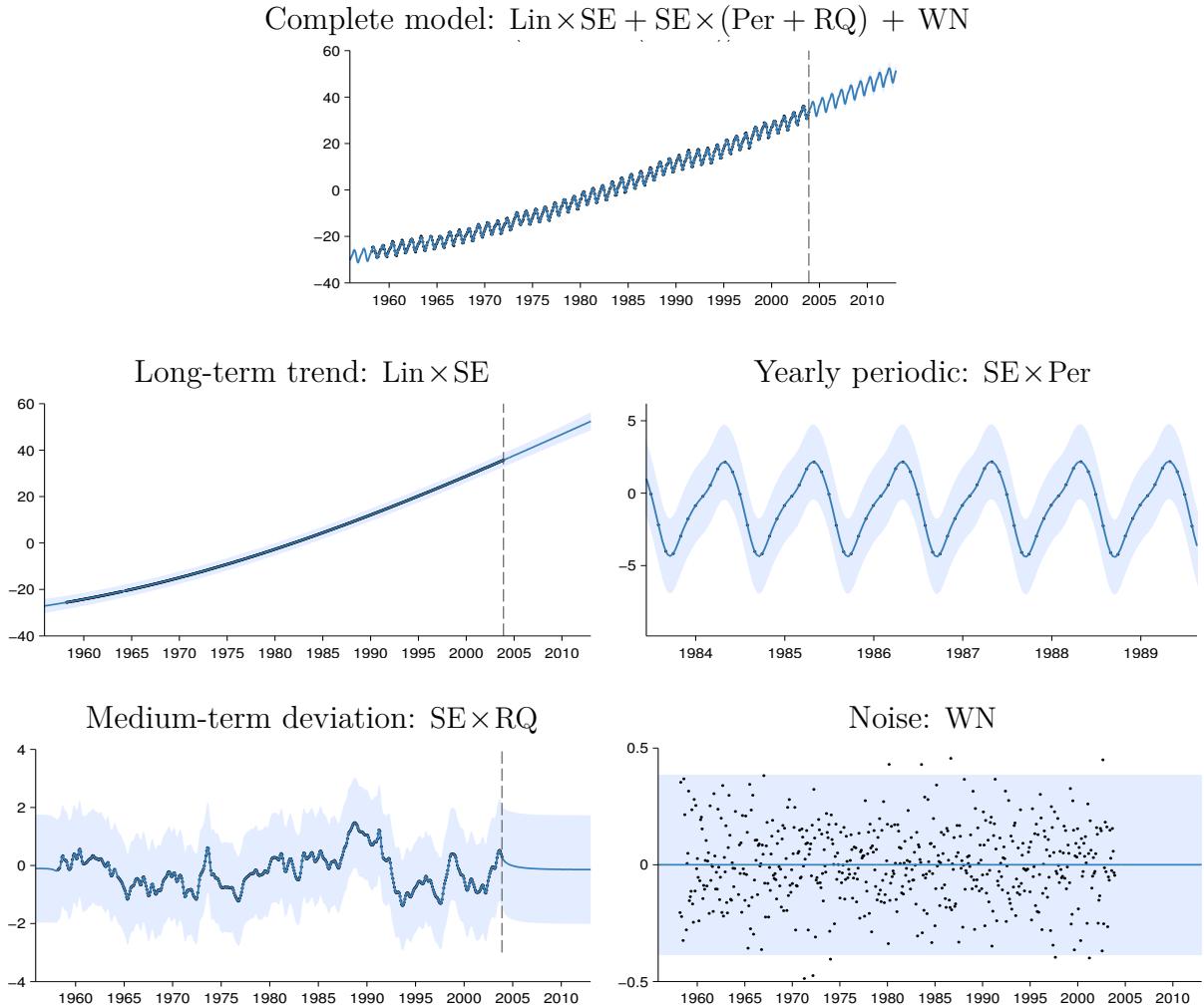


Figure 3.4 *First row*: The full posterior on the Mauna Loa dataset, after a search of depth 10. *Subsequent rows*: The automatic decomposition of the time series. The model is a sum of long-term, yearly periodic, medium-term components, and residual noise, respectively. The yearly periodic component has been rescaled for clarity.

slowly-changing structure while others capture quickly-varying structure, but there is no hard distinction between “signal” components and “noise” components.

### 3.6.2 Airline passenger counts

Figure 3.5 shows the decomposition produced by applying our method to monthly totals of international airline passengers (Box et al., 1970). We observe similar components to those in the Mauna Loa dataset: a long term trend, annual periodicity, and medium-term deviations. In addition, the composite kernel captures the near-linearity of the

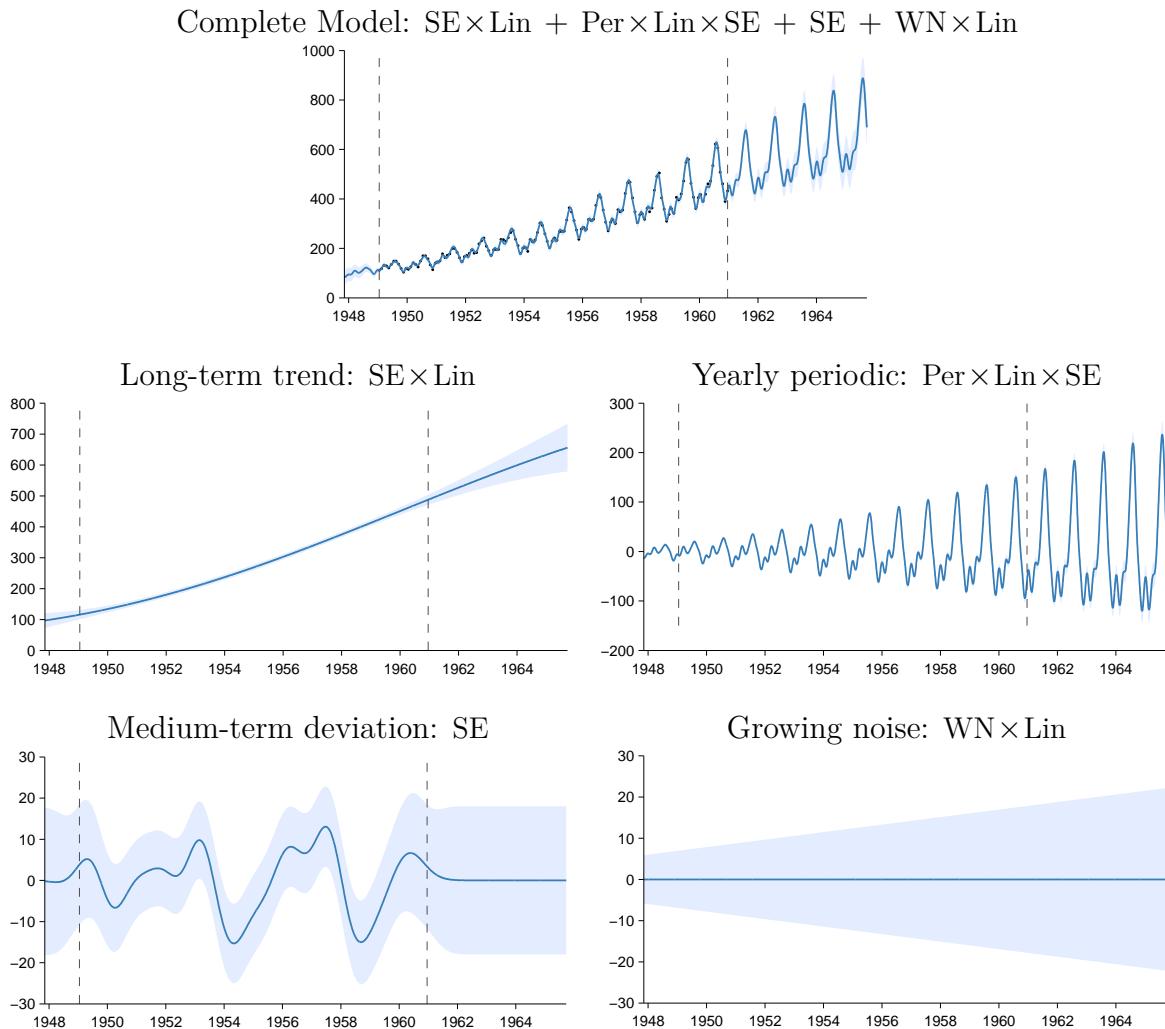


Figure 3.5 *First row:* The airline dataset and posterior after a search of depth 10. *Subsequent rows:* Additive decomposition of posterior into long-term smooth trend, yearly variation, and short-term deviations. Due to the linear kernel, the marginal variance grows over time, making this a heteroskedastic model.

long-term trend, and the linearly growing amplitude of the annual oscillations.

## 3.7 Related work

### Building kernel functions by hand

Rasmussen and Williams (2006, Chapter 5) devote 4 pages to manually constructing a composite kernel to model the Mauna Loa dataset. Other examples of papers whose

main contribution is to manually construct and fit a composite GP kernel are [Klenske et al. \(2013\)](#); [Lloyd \(2013\)](#); [Preotiuc-Pietro and Cohn \(2013\)](#).

## Nonparametric regression in high dimensions

Nonparametric regression methods such as splines, locally-weighted regression, and GP regression are popular because they are capable of learning arbitrary smooth functions of the data. Unfortunately, they suffer from the curse of dimensionality: it is very difficult for these models to generalize well in more than a few dimensions.

Applying nonparametric methods in high-dimensional spaces can require imposing additional structure on the model. One such structure is additivity. Generalized additive models ([Hastie and Tibshirani, 1990](#)) assume the regression function is a transformed sum of functions defined on the individual dimensions:  $\mathbb{E}[f(\mathbf{x})] = g^{-1}(\sum_{d=1}^D f_d(x_d))$ . These models have a limited compositional form, but one which is interpretable and often generalizes well. In our grammar, we can capture such structure through sums of base kernels along different dimensions, although we have not yet tried incorporating a warping function  $g(\cdot)$ .

It is possible to add more flexibility to additive models by considering higher-order interactions between different dimensions. Chapter 6 considers GP models whose kernel implicitly sums over all possible interactions of input variables. [Plate \(1999\)](#) constructs a special case of this model class, summing an SE kernel along each dimension, with a single SE-ARD kernel (a product of SE over all dimensions). Both of these models can be expressed in our grammar.

A closely related procedure is smoothing-splines ANOVA ([Gu, 2002](#); [Wahba, 1990](#)). This model is a linear combinations of splines along each dimension, all pairs of dimensions, and possibly higher-order combinations. Because the number of terms to consider grows exponentially in the order, in practice, only terms of first and second order are usually considered.

Semi-parametric regression (e.g. [Ruppert et al., 2003](#)) attempts to combine interpretability with flexibility by building a composite model out of an interpretable, parametric part (such as linear regression) and a ‘catch-all’ nonparametric part (such as a GP with an SE kernel). This model class can be represented through the kernel SE + Lin.

## Kernel learning

There is a large body of work attempting to construct a rich kernel through a weighted sum of many base kernels (e.g. Bach, 2009a; Gönen and Alpaydin, 2011). These approaches usually admit a convex objective function. However the component kernels, as well as their parameters, must be specified in advance.

Salakhutdinov and Hinton (2008) use a deep neural network with unsupervised pre-training to learn an embedding  $g(\mathbf{x})$  onto which a GP with an SE kernel is placed:  $\text{Cov}[f(\mathbf{x}), f(\mathbf{x}')] = k(g(\mathbf{x}), g(\mathbf{x}'))$ . This is a flexible approach to kernel learning, but relies upon finding structure in the input density  $p(\mathbf{x})$ . Instead, we focus on domains where most of the interesting structure is in  $f(\mathbf{x})$ .

Sparse spectrum GPs (Lázaro-Gredilla et al., 2010) approximate the spectral density of a stationary kernel function using delta functions which corresponds to kernels of the form  $\sum \cos$ . Similarly, Wilson and Adams (2013) introduce spectral mixture kernels, which approximate the spectral density using a mixture of Gaussians – corresponding to kernels of the form  $\sum \text{SE} \times \cos$ . Both groups demonstrate, using Bochner’s theorem (Bochner, 1959), that these kernels can approximate any stationary covariance function. Our language of kernels includes both of these kernel classes (see table 3.1).

There is a large body of work attempting to construct rich kernels through a weighted sum of base kernels called multiple kernel learning (MKL) (e.g. Bach et al., 2004). These approaches find the optimal solution in polynomial time, but only if the component kernels and parameters are pre-specified. We compare to a Bayesian variant of MKL in section 3.8, expressed as a restriction of our language of kernels.

## Equation learning

Todorovski and Džeroski (1997), Washio et al. (1999) and Schmidt and Lipson (2009) learn parametric forms of functions specifying time series, or relations between quantities. In contrast, ABCD learns a parametric form for the covariance, allowing it to model functions which do not have a simple parametric form but still have high-level structure. An examination of the structure discovered by the automatic equation-learning software Eureqa (Schmidt and Lipson, 2013) on the airline and Mauna Loa datasets can be found in Lloyd et al. (2014).

## Structure discovery through grammars

Kemp and Tenenbaum (2008) learned the structural form of a graph used to model

human similarity judgements. Their grammar on graph structures includes planes, trees, and cylinders. Some of their discrete graph structures have continuous analogues in our own space. For example,  $\text{SE}_1 \times \text{SE}_2$  and  $\text{SE}_1 \times \text{Per}_2$  can be seen as mapping the data to a plane and a cylinder, respectively. Section 2.8 examines these structures in more detail.

Diosan et al. (2007) and Bing et al. (2010) learn composite kernels for support vector machines and relevance vector machines, respectively, using genetic search algorithms to optimize cross-validation error. Similarly, Kronberger and Kommenda (2013) search over composite kernels for GPs using genetic programming, optimizing the unpenalized marginal likelihood. These methods explore similar languages of kernels to the one explored in this chapter. It is not clear whether the complex genetic searches used by these methods offer advantages over the straightforward but naïve greedy search used in this chapter. Our search criterion has the advantages of being both differentiable with respect to kernel parameters, and trading off model fit and complexity automatically. This prior work also did not explore the automatic model decomposition, summarization and description made possible by the use of GP models.

Grosse et al. (2012) performed a greedy search over a compositional model class for unsupervised learning, using a grammar of matrix decomposition models, and a greedy search procedure based on held-out likelihood. This model class contains many existing unsupervised models as special cases, and was able to discover such structure automatically from data. Our framework takes a similar approach, but in a supervised setting.

Similarly, Steinruecken (2014) showed to automatically perform inference in arbitrary compositions of discrete sequence models. More generally, Dechter et al. (2013) and Liang et al. (2010) constructed grammars over programs, and automatically searched the resulting spaces.

## 3.8 Experiments

### 3.8.1 Interpretability versus accuracy

BIC trades off model fit and complexity by penalizing the number of parameters in a kernel expression. This can result in ABCD favoring kernel expressions with nested products of sums, producing descriptions involving many additive components when expressions are expanded. While these models typically have good predictive performance, the large number of components can make them less interpretable. We experimented

with not allowing parentheses during the search, discouraging nested expressions. This was achieved by distributing all products immediately after each search operator was applied. We call this procedure ABCD-interpretability, in contrast to the unrestricted version of the search, ABCD-accuracy.

### 3.8.2 Predictive accuracy on time series

We evaluate the performance of the algorithms listed below on 13 real time-series from various domains from the time series data library ([Hyndman](#), accessed July 2013).

#### Algorithms

We compare ABCD to equation learning using Eureqa ([Schmidt and Lipson, 2013](#)), as well as six other regression algorithms: linear regression, GP regression with a single SE kernel (squared exponential), a Bayesian variant of multiple kernel learning (MKL) (e.g. [Bach et al., 2004](#); [Gönen and Alpaydin, 2011](#)), changepoint modeling (e.g. [Fox and Dunson, 2013](#); [Garnett et al., 2010](#); [Saatçi et al., 2010](#)), spectral mixture kernels ([Wilson and Adams, 2013](#)) (spectral kernels), and trend-cyclical-irregular models (e.g. [Lind et al., 2006](#)).

We set Eureqa’s search objective to the default mean-absolute-error. All algorithms besides Eureqa can be expressed as restrictions of our modeling language (see table 3.1), so we perform inference using the same search and objective function, with appropriate restrictions to the language. For MKL, trend-cyclical-irregular, and spectral kernels, the greedy search procedure of ABCD corresponds to a forward-selection algorithm. For squared-exponential and linear regression, the procedure corresponds to standard marginal likelihood optimization.

We restricted to regression algorithms for comparability; we did not include models which regress on previous values of times series, such as auto-regressive or moving-average models (e.g. [Box et al., 1970](#)). Constructing a language of autoregressive time-series models would be an interesting area for future research.

#### Extrapolation experiments

To test extrapolation, we trained all algorithms on the first 90% of the data, predicted the remaining 10% and then computed the root mean squared error (RMSE). The RMSEs are then standardised by dividing by the smallest RMSE for each data set, so the best performance on each data set will have a value of 1.

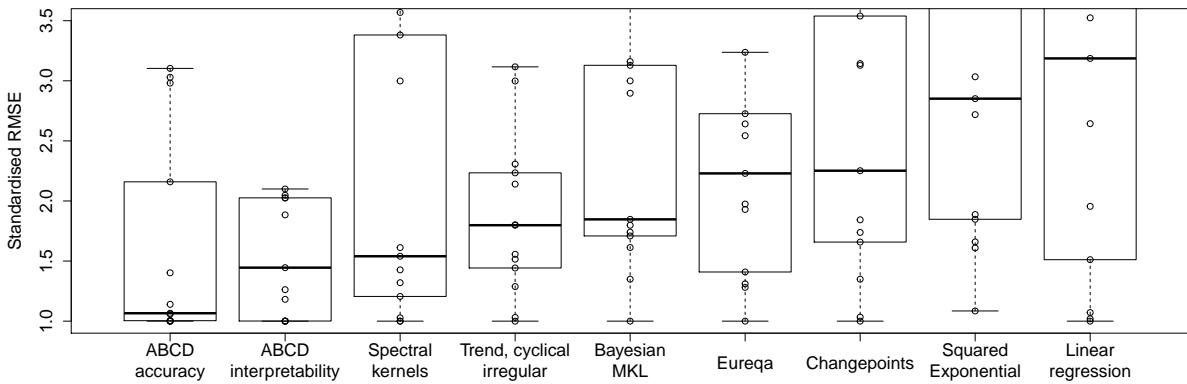


Figure 3.6 Box plot (showing median and quartiles) of standardised extrapolation RMSE (best performance = 1) on 13 time-series. The methods are ordered by median.

Figure 3.6 shows the standardised RMSEs across algorithms. ABCD-accuracy outperforms ABCD-interpretability. Both algorithms have lower quartiles than all other methods.

Overall, the model construction methods with greater capacity perform better: ABCD outperforms trend-cyclical-irregular, which outperforms Bayesian MKL, which outperforms squared-exponential. Despite searching over a rich model class, Eureqa performs relatively poorly, since very few datasets are parsimoniously explained by a parametric equation.

Not shown on the plot are large outliers for spectral kernels, Eureqa, squared exponential and linear regression with normalized RMSEs of 11, 493, 22 and 29 respectively.

### 3.8.3 High-dimensional prediction

ABCD can also be applied to multidimensional regression problems without modification. An experimental comparison with other methods can be found in section 6.6, where it outperforms a wide variety of multidimensional regression methods.

### 3.8.4 Structure recovery on synthetic data

The structure found in the examples above may seem reasonable, but we may wonder to what extent ABCD is consistent - that is, when do we recover all the structure in any given dataset? It is difficult to tell from predictive accuracy alone if the search procedure is finding the correct structure, especially in multiple dimensions. To address

this question, we tested our method’s ability to recover known structure on a set of synthetic datasets.

For several composite kernel expressions, we constructed synthetic data by first sampling 300 points uniformly at random, then sampling function values at those points from a GP prior. We then added i.i.d. Gaussian noise to the functions at various signal-to-noise ratios (SNR).

Table 3.2 Kernels chosen by our method on synthetic data generated using known kernel structures.  $D$  denotes the dimension of the functions being modeled. SNR indicates the signal-to-noise ratio. Dashes (–) indicate no structure was found. Each kernel implicitly has a WN kernel included.

True kernel	$D$	SNR = 10	SNR = 1	SNR = 0.1
SE + RQ	1	SE	SE × Per	SE
Lin × Per	1	Lin × Per	Lin × Per	SE
SE <sub>1</sub> + RQ <sub>2</sub>	2	SE <sub>1</sub> + SE <sub>2</sub>	Lin <sub>1</sub> + SE <sub>2</sub>	Lin <sub>1</sub>
SE <sub>1</sub> + SE <sub>2</sub> × Per <sub>1</sub> + SE <sub>3</sub>	3	SE <sub>1</sub> + SE <sub>2</sub> × Per <sub>1</sub> + SE <sub>3</sub>	SE <sub>2</sub> × Per <sub>1</sub> + SE <sub>3</sub>	–
SE <sub>1</sub> × SE <sub>2</sub>	4	SE <sub>1</sub> × SE <sub>2</sub>	Lin <sub>1</sub> × SE <sub>2</sub>	Lin <sub>2</sub>
SE <sub>1</sub> × SE <sub>2</sub> + SE <sub>2</sub> × SE <sub>3</sub>	4	SE <sub>1</sub> × SE <sub>2</sub> + SE <sub>2</sub> × SE <sub>3</sub>	SE <sub>1</sub> + SE <sub>2</sub> × SE <sub>3</sub>	SE <sub>1</sub>
(SE <sub>1</sub> + SE <sub>2</sub> ) × (SE <sub>3</sub> + SE <sub>4</sub> )	4	(SE <sub>1</sub> + SE <sub>2</sub> ) × ... (SE <sub>3</sub> × Lin <sub>3</sub> × Lin <sub>1</sub> + SE <sub>4</sub> )	(SE <sub>1</sub> + SE <sub>2</sub> ) × ... SE <sub>3</sub> × SE <sub>4</sub>	–

Table 3.2 shows the results. For the highest SNR, the method finds all relevant structure except in one case. The reported additional linear structure in the last row is explainable by the fact that functions sampled from SE kernels with long length-scales occasionally have near-linear trends. As the noise increases, our method generally backs off to simpler structures, rather than over-fitting.

## Source code

Source code to perform all experiments is available at [github.com/jamesrobertlloyd/gpss-research](https://github.com/jamesrobertlloyd/gpss-research). All GP parameter optimisation was performed by automated calls to the GPML toolbox, available at [gaussianprocess.org/gpml/code](http://gaussianprocess.org/gpml/code).

### 3.9 Discussion

Towards the goal of automating statistical modeling, we developed a system which constructs an appropriate model from an open-ended language, and automatically generates plots decomposing the different types of structure present in the model.

This was done by introducing a space of kernels defined by sums and products of a small number of base kernels. The set of models in this space includes many standard regression models. We proposed a search procedure for this space of kernels, and argued that this search process parallels the process of scientific discovery, and of model-building by statisticians.

We found that the learned structures are often capable of accurate extrapolation in time-series datasets, and are competitive with widely used kernel classes and kernel combination methods on a variety of prediction tasks. The learned kernels often yield decompositions of a signal into diverse and interpretable components, enabling model-checking by humans. We hope that this procedure has the potential to make powerful statistical model-building techniques accessible to non-experts.

The next chapter will show how the model components found by this procedure can be automatically described using English-language text.

# Chapter 4

## Automatic Model Description

“When I write, I feel like an armless, legless man with a crayon in his mouth.”

– *Kurt Vonnegut*

The previous chapter showed how to automatically build structured models through a language of kernels. It also showed how to decompose the resulting models into the different types of structure present, and how to visually illustrate the type of structure captured by each component. This chapter shows how automatically describe the resulting model structures with English text.

The main idea is to treat every component of a product kernel as an adjective, or as a short phrase which modifies the description of a kernel. To see how this could work, recall that the model decomposition plots of chapter 3 showed that most of the structure in each component was determined by that component’s kernel. Even across different datasets, the meaning of individual parts of kernels is consistent in some ways. For example, Per indicates repeating structure, and SE indicates smooth change over time.

This chapter also presents a system which combines automatically generated text and plots to generate reports which highlight interpretable features discovered in a variety of data sets. An example of a complete automatically-generated report can be found in chapter D.

The work appearing in this chapter was written in collaboration with James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani, and was published in [Lloyd et al. \(2014\)](#). The procedure translating kernels into adjectives grew out of discussions between James and myself. James Lloyd wrote the code to automatically generate reports, and ran all of the experiments.

## 4.1 Generating descriptions of composite kernels

There are two main features of our language of GP models that allow description to be performed automatically. First, the sometimes complicated kernel expressions found by model search can always be simplified into a sum of products. As discussed in chapter 2, a sum of kernels corresponds to a sum of functions, so each product can be described separately. Second, each kernel in a product modifies the resulting model in a consistent way. Therefore, one can describe a product of kernels by concatenating descriptions of the effect of each part of the product. One part needs to be described using a noun.

For example, we can describe the product of kernels  $\text{Per} \times \text{SE}$  by representing  $\text{Per}$  by a noun (“a periodic function”) modified by a phrase representing the effect of the  $\text{SE}$  kernel (“whose shape varies smoothly over time”). To simplify the system, we restrict the base kernels to the set  $\{\text{C}, \text{Lin}, \text{WN}, \text{SE}, \text{Per}, \text{and } \boldsymbol{\sigma}\}$ . Recall that the sigmoidal kernel  $\boldsymbol{\sigma}(x, x') = \sigma(x)\sigma(x')$  allows changepoints and change-windows.

### 4.1.1 Simplification rules

In order to be able to use the same phrase to describe the effect of each base kernel in different circumstances, we must convert each kernel expression into a standard, simplified form.

First, we distribute all products of sums into sums of products. Then, we apply several simplification rules to the kernel expression:

- Products of two or more  $\text{SE}$  kernels can be equivalently represented by a single  $\text{SE}$  with different parameters.
- Multiplying the white-noise kernel ( $\text{WN}$ ) by any stationary kernel ( $\text{C}$ ,  $\text{WN}$ ,  $\text{SE}$ , or  $\text{Per}$ ) gives another  $\text{WN}$  kernel.
- Multiplying any kernel by the constant kernel ( $\text{C}$ ) only changes the parameters of the original kernel, and so can be factored out of any product in which it appears.

After applying these rules, any composite kernel can be written as a sum of terms of the form:

$$K \prod_m \text{Lin}^{(m)} \prod_n \boldsymbol{\sigma}^{(n)}, \quad (4.1)$$

where  $K$  is one of  $\{\text{WN}, \text{C}, \text{SE}, \prod_k \text{Per}^{(k)}\}$  or  $\{\text{SE} \prod_k \text{Per}^{(k)}\}$ , and  $\prod_i k^{(i)}$  denotes a product of kernels, each with different parameters. Superscripts denote different instances of the same kernel appearing in a product:  $\text{SE}^{(1)}$  can have different kernel parameters than  $\text{SE}^{(2)}$ .

### 4.1.2 Describing each part of a product of kernels

Loosely speaking, each kernel in a product modifies the resulting GP model in a consistent way. This allows us to describe the contribution of each kernel in a product as an adjective, or more generally as a post-modifier of a noun. We now describe how each of the kernels in our grammar modifies a GP model:

- **Multiplication by SE** removes long range correlations from a model, since  $\text{SE}(x, x')$  decreases monotonically to 0 as  $|x - x'|$  increases. This will convert any global correlation structure into local correlation only.
- **Multiplication by Lin** is equivalent to multiplying the function being modeled by a linear function. If  $f(x) \sim \text{GP}(0, k)$ , then  $xf(x) \sim \text{GP}(0, \text{Lin} \times k)$ . This causes the standard deviation of the model to vary linearly, without affecting the correlation between function values.
- **Multiplication by  $\sigma$**  is equivalent to multiplying the function being modeled by a sigmoid, which means that the function goes to zero before or after some point.
- **Multiplication by Per** removes correlations between all pairs of function values not close to one period apart.
- **Multiplication by any kernel** modifies the covariance in the same way as multiplying by a function drawn from a corresponding GP prior. This follows from the fact that if  $f_1(x) \sim \text{GP}(0, k_1)$  and  $f_2(x) \sim \text{GP}(0, k_2)$  then

$$\text{Cov}[f_1(x)f_2(x), f_1(x')f_2(x')] = k_1(x, x')k_2(x, x'). \quad (4.2)$$

Put more plainly, a GP whose covariance is a product of kernels has the same covariance as a product of two functions, each drawn from the corresponding GP prior. However, the distribution of  $f_1f_2$  is not GP distributed – it has higher non-central moments as well. This identity can be used to generate a cumbersome “worst-case” description in cases where a more concise description of the effect of a

kernel is not known. For example, it is used in our system to describe the product of more than one periodic function.

Table 4.1 gives the corresponding description of the effect of each type of kernel in a product, written as a post-modifier.

Kernel	Postmodifier phrase
SE	whose shape changes smoothly
Per	modulated by a periodic function
Lin	with linearly varying amplitude
$\prod_k \text{Lin}^{(k)}$	with polynomially varying amplitude
$\prod_k \sigma^{(k)}$	which applies until / from [changepoint]

Table 4.1 Descriptions of the effect of each kernel, written as a post-modifier.

Table 4.2 gives the corresponding description of each kernel before it has been multiplied by any other, written as a noun phrase.

Kernel	Noun phrase
WN	uncorrelated noise
C	constant
SE	smooth function
Per	periodic function
Lin	linear function
$\prod_k \text{Lin}^{(k)}$	polynomial

Table 4.2 Noun phrase descriptions of each type of kernel

### 4.1.3 Combining descriptions into noun phrases

In order to build a noun phrase describing a product of kernels, we choose one kernel to act as a head noun which is modified by appending descriptions of the other kernels in the product.

As an example, a kernel of the form  $\text{Per} \times \text{Lin} \times \sigma$  could be described as a

$\underbrace{\text{Per}}_{\text{periodic function}} \quad \times \quad \underbrace{\text{Lin}}_{\text{with linearly varying amplitude}} \quad \times \quad \underbrace{\sigma}_{\text{which applies until 1700.}}$

where Per was chosen to be the head noun.

In our system, the head noun is chosen according to the following ordering:

$$\text{Per, WN, SE, C, } \prod_m \text{Lin}^{(m)}, \prod_n \boldsymbol{\sigma}^{(n)} \quad (4.3)$$

Combining tables 4.1 and 4.2 with ordering 4.3 provides a general method to produce descriptions of kernels.

### Extensions

In practice, we also incorporate a number of other rules which help to make the descriptions shorter, easier to parse, or clearer:

- We add extra adjectives depending on kernel parameters. For example, an SE with a relatively short lengthscale might be described as “a rapidly-varying smooth function” as opposed to just “a smooth function”.
- Descriptions can include kernel parameters. For example, we might write that a function is “repeating with a period of 7 days”.
- Descriptions can include extra information about the model not contained in the kernel. For example, based on the slope of the posterior mean, we might “a linearly increasing function” as opposed to “a linear function”.
- Some kernels can be described through pre-modifiers. For example, we might write “an approximately periodic function” as opposed to “a periodic function whose shape changes smoothly”.

### Ordering additive components

The reports generated by our system attempt to present the most interesting or important features of a data set first. As a heuristic, we order components by always adding next the component which most reduces the 10-fold cross-validated mean absolute error.

#### 4.1.4 Worked example

This section shows an example of the description procedure, starting with a compound kernel containing every type of base kernel in our set:

$$\text{SE} \times (\text{WN} \times \text{Lin} + \text{CP(C, Per))). \quad (4.4)$$

This is converted to a sum of products, and the changepoint is converted into sigmoidal kernels (recall the definition of changepoint kernels in section 2.5:

$$\text{SE} \times \text{WN} \times \text{Lin} + \text{SE} \times \text{C} \times \boldsymbol{\sigma} + \text{SE} \times \text{Per} \times \bar{\boldsymbol{\sigma}}. \quad (4.5)$$

which is simplified using the rules in section 4.1.1 to

$$\text{WN} \times \text{Lin} + \text{SE} \times \boldsymbol{\sigma} + \text{SE} \times \text{Per} \times \bar{\boldsymbol{\sigma}}. \quad (4.6)$$

To describe the first component, ( $\text{WN} \times \text{Lin}$ ), the head noun description for WN, “uncorrelated noise”, is concatenated with a modifier for Lin, “with linearly increasing standard deviation”.

The second component, ( $\text{SE} \times \boldsymbol{\sigma}$ ), is described as “A smooth function with a length-scale of [lengthscale] [units]”, corresponding to the SE, “which applies until [changepoint]”.

Finally, the third component,  $\text{SE} \times \text{Per} \times \bar{\boldsymbol{\sigma}}$ , is described as “An approximately periodic function with a period of [period] [units] which applies from [changepoint]”.

## 4.2 Example descriptions

In this section, we demonstrate the ability of our procedure, ABCD, to write intelligible descriptions of the structure present in two time series. The model descriptions were generated based off of models produced by the automatic search method presented in chapter 3.

### 4.2.1 Summarizing 400 years of solar activity

First, we show excerpts from the report automatically generated on annual solar irradiation data from 1610 to 2011. This dataset is shown in figure 4.1.

This time series has two pertinent features: First, a roughly 11-year cycle of solar activity. Second, a period lasting from 1645 to 1715 having almost no variance. This flat region corresponds to the Maunder minimum, a period in which sunspots were extremely rare (Lean et al., 1995). The Maunder minimum is an example of the type of structure which can be captured by changewindows.

The first section of each report generated by ABCD is a summary of the structure found in the dataset. Figure 4.2 show natural-language summaries of the top four com-

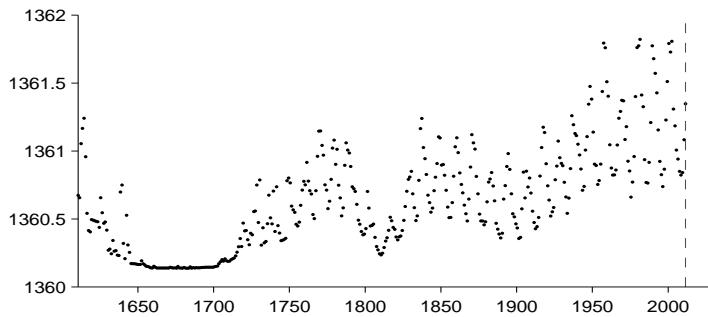


Figure 4.1 Solar irradiance data.

- A constant.
- A constant. This function applies from 1643 until 1716.
- A smooth function. This function applies until 1643 and from 1716 onwards.
- An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards.

Figure 4.2 Automatically generated descriptions of the first four components discovered by ABCD on the solar irradiance data set. The dataset has been decomposed into diverse structures with simple descriptions.

ponents discovered by ABCD on the solar dataset. From these short summaries, we can see that the system has identified the Maunder minimum (second component) and the 11-year solar cycle (fourth component). These components are visualized and described in figures 4.3 and 4.5, respectively. The third component, visualized in figure 4.4, captures the smooth variation over time of the overall level of solar activity.

This component is constant. This component applies from 1643 until 1716.

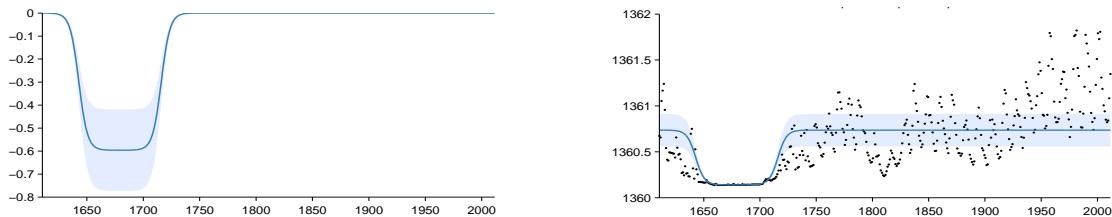


Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Figure 4.3 Extract from an automatically-generated report describing the model component corresponding to the Maunder minimum.

This component is a smooth function with a typical lengthscale of 23.1 years. This component applies until 1643 and from 1716 onwards.

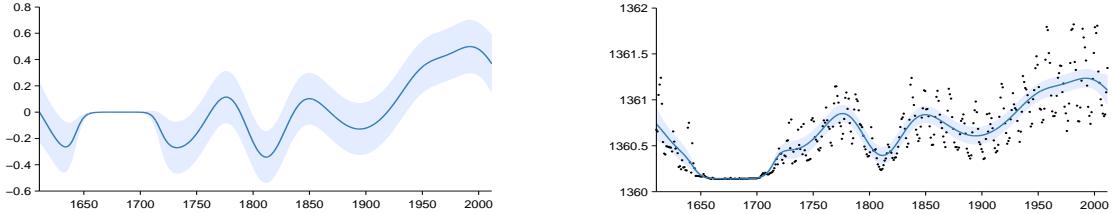


Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

Figure 4.4 Characterizing the medium-term smoothness of solar activity levels. By allowing other components to explain the periodicity, noise, and the Maunder minimum, ABCD can isolate the part of the signal best explained by a slowly-varying trend.

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.

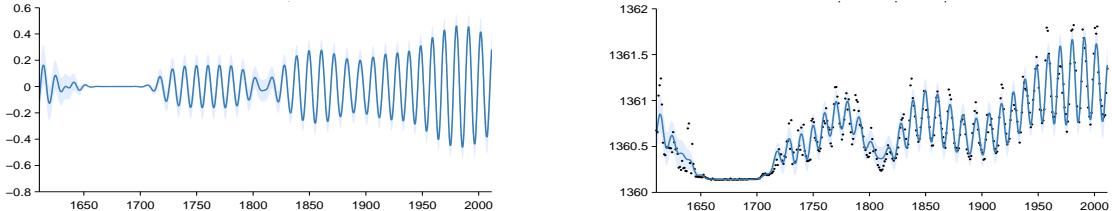


Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

Figure 4.5 This part of the report isolates and describes the approximately 11-year sunspot cycle, also noting its disappearing during the Maunder minimum.

The complete report generated on this dataset can be found in chapter D. This report also contains samples from the model posterior.

#### 4.2.2 Describing changing noise levels

Next, we present excerpts of the description generated by our procedure on international airline passenger counts over time. The dataset and model described is the same as that

shown in figure 3.5. High-level descriptions of the four components discovered are shown in figure 4.6.

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.

Figure 4.6 Short descriptions of the four components of the airline model.

This component is approximately periodic with a period of 1.0 years and varying amplitude. Across periods the shape of this function varies very smoothly. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.

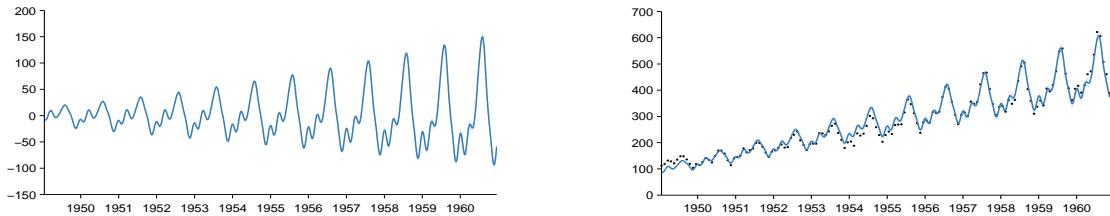


Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Figure 4.7 Describing non-stationary periodicity in the airline data.

This component models uncorrelated noise. The standard deviation of the noise increases linearly.

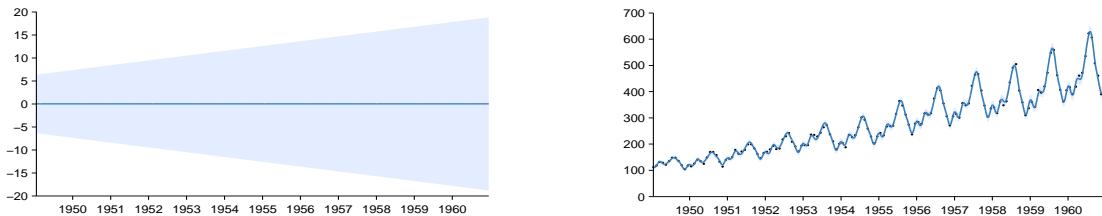


Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

Figure 4.8 Modeling time-changing variance in the airline dataset.

The second component, shown in figure 4.7, is accurately described as approximately (SE) periodic (Per) with linearly growing amplitude (Lin).

The description of the fourth component, shown in figure 4.8, expresses the fact that

the scale of the unstructured noise in the model grows linearly with time.

The complete report generated on this dataset can be found in the supplementary material of Lloyd et al. (2014). Other example reports describing a wide variety of time-series can be found at [mlg.eng.cam.ac.uk/lloyd/abcdoutput/](http://mlg.eng.cam.ac.uk/lloyd/abcdoutput/)

### Source code

Source code to perform all experiments is available at [github.com/jamesrobertlloyd/gpss-research](https://github.com/jamesrobertlloyd/gpss-research).

## 4.3 Related work

To the best of our knowledge, our procedure is the first example of automatic description of nonparametric statistical models. However, systems with natural language output have been built in the areas of video interpretation (Barbu et al., 2012) and automated theorem proving (Ganesalingam and Gowers, 2013).

## 4.4 Limitations of this approach

This chapter presents a procedure for describing GP models which relied on the particular properties of the model and kernel being used. The compositional nature of the descriptions avoided an exponential blowup in the number of special cases that needed to be considered. However, during development, we noted several difficulties with this approach:

- **Some kernels are hard to describe.** For instance, we did not include the RQ kernel in the text-generation procedure. This was done for several reasons. First, the RQ kernel can be equivalently expressed as a scale mixture of SE kernels. Second, it was difficult to think of a description for the hyperparameter which controls the heaviness of the tails of the RQ kernel. Third, a product of two RQ kernels does not give another RQ kernel, which raises the question of how to concisely describe products of RQ kernels.
- **Reliance on additivity.** Much of the work of decomposing models was due to the additive decomposition. However, it is easy to imagine data without such additive decomposition, or which only decomposes additively after being warped.

Describing the properties of transformations of the data may not be as useful to the end user.

## 4.5 Conclusions

Towards the goal of automating statistical modeling we presented a system which, given a structured GP model, automatically generates detailed reports that describe patterns in the data captured by the model. This, combined with the automatic model search of chapter 3, gives a procedure which can discover and describe a variety of patterns in time series.

# Chapter 5

## Deep Gaussian Processes

“I asked myself: On any given day, would I rather be wrestling with a sampler, or proving theorems?”

– Peter Orbanz, personal communication

Choosing appropriate architectures and regularization strategies of deep networks is important for good predictive performance. In this chapter, we propose to study this problem by viewing deep nets as priors on functions. By viewing neural networks this way, we can analyze their properties without reference to any particular dataset, loss function, or training method. Instead, we can ask what sorts of information-processing structures these priors give rise to, and check whether those structures are the same sort which we expect to find in useful models.

As a starting point, we will relate neural networks to Gaussian processes, and examine a class of infinitely-wide, deep neural networks called *deep Gaussian processes*. Deep GPs are an attractive model class to study for several reasons. [Damianou and Lawrence \(2013\)](#) showed that the probabilistic nature of deep GPs guards against overfitting, and allows us to use the marginal likelihood to automatically tune the model architecture without the need for cross-validation. [Hensman et al. \(2014a\)](#) showed that stochastic variational inference is possible in deep GPs. Together, these results suggest that deep GPs are a promising alternative to neural nets. For the analysis in this chapter, deep GPs are attractive because they abstract away some of the details of finite neural networks.

Our analysis will show that in standard architectures, the representational capacity of standard deep networks tends to decrease as the number of layers increases, retaining only a single degree of freedom in the limit. We propose an alternate network architecture that connects the input to each layer that does not suffer from this pathology. We also examine *deep kernels*, obtained by composing arbitrarily many fixed feature transforms.

The ideas contained in this chapter were developed through discussions with Oren Rippel, Ryan Adams and Zoubin Ghahramani, and appear in [Duvenaud et al. \(2014\)](#).

## 5.1 Definition and relation to neural networks

In this chapter, we'll define a deep GP as any prior on functions constructed by composing draws from GP priors. An example of a deep GP is a composition of vector-valued functions, where each function in each layer is drawn independently from a GP prior:

$$\begin{aligned} \mathbf{f}^{(1:L)}(\mathbf{x}) &= \mathbf{f}^{(L)}(\mathbf{f}^{(L-1)}(\dots \mathbf{f}^{(2)}(\mathbf{f}^{(1)}(\mathbf{x})) \dots)) \\ \text{with each } f_d^{(\ell)} &\stackrel{\text{ind}}{\sim} \text{GP}\left(0, k_d^\ell(\mathbf{x}, \mathbf{x}')\right) \end{aligned} \quad (5.1)$$

A multilayer neural network also implements a composition of vector-valued functions, one per layer. The rest of this section shows the precise relationship between Gaussian processes and neural nets, as well as two equivalent ways of constructing neural networks which give rise to deep GPs.

### 5.1.1 Single-layer models

First, we will relate neural networks with one hidden layer to Gaussian processes, using the standard neural network architecture known as the multi-layer perceptron (MLP). In the typical definition of an MLP, the hidden units of the first layer are defined as:

$$\mathbf{h}^{(1)}(\mathbf{x}) = \sigma(\mathbf{b}^{(1)} + \mathbf{W}^{(1)}\mathbf{x}) \quad (5.2)$$

where  $\mathbf{h}$  are the hidden unit activations,  $\mathbf{b}$  is a bias vector,  $\mathbf{W}$  is a weight matrix and  $\sigma$  is a one-dimensional nonlinear function applied element-wise. The output vector  $f(\mathbf{x})$  is simply a weighted sum of these hidden unit activations:

$$f(\mathbf{x}) = \mathbf{V}^{(1)}\sigma(\mathbf{b}^{(1)} + \mathbf{W}^{(1)}\mathbf{x}) = \mathbf{V}^{(1)}\mathbf{h}^{(1)}(\mathbf{x}) \quad (5.3)$$

where  $\mathbf{V}^{(1)}$  is another weight matrix.

[Neal \(1995, chapter 2\)](#) showed that GPs can be viewed as a prior on neural networks with infinitely many hidden units and unknown weights. More precisely, for any model

of the form

$$f(\mathbf{x}) = \frac{1}{K} \boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^K \alpha_i h_i(\mathbf{x}), \quad (5.4)$$

with fixed<sup>1</sup> features  $[h_1(\mathbf{x}), \dots, h_K(\mathbf{x})]^\top = \mathbf{h}(\mathbf{x})$  and i.i.d.  $\alpha$ 's with zero mean and finite variance  $\sigma^2$ , the central limit theorem implies that as the number of features  $K$  grows, any two function values  $f(\mathbf{x}), f(\mathbf{x}')$  have a joint distribution approaching a Gaussian:

$$\lim_{K \rightarrow \infty} p\left(\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}') \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{\sigma^2}{K} \begin{bmatrix} \sum_{i=1}^K h_i(\mathbf{x})h_i(\mathbf{x}) & \sum_{i=1}^K h_i(\mathbf{x})h_i(\mathbf{x}') \\ \sum_{i=1}^K h_i(\mathbf{x}')h_i(\mathbf{x}) & \sum_{i=1}^K h_i(\mathbf{x}')h_i(\mathbf{x}') \end{bmatrix}\right) \quad (5.5)$$

A joint Gaussian distribution between any set of function values is the definition of a Gaussian process.

The result is surprisingly general: it puts no constraints on the features (other than having uniformly bounded activation), nor does it require that the feature weights  $\boldsymbol{\alpha}$  be Gaussian distributed.

We can also work backwards to derive a one-layer MLP from any GP. Mercer's theorem implies that any positive-definite kernel function corresponds to an inner product of features:  $k(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}')$ . Thus in the one-hidden-layer case, the correspondence between MLPs and GPs is simple: the implicit features  $\mathbf{h}(\mathbf{x})$  of the kernel correspond to the hidden units of the MLP.

### 5.1.2 Multiple hidden layers

In an MLP with multiple hidden layers, activation of the  $\ell$ th layer units are given by

$$\mathbf{h}^{(\ell)}(\mathbf{x}) = \sigma \left( \mathbf{b}^{(\ell)} + \mathbf{W}^{(\ell)} \mathbf{h}^{(\ell-1)}(\mathbf{x}) \right). \quad (5.6)$$

This architecture is shown on the right of figure 5.1. For example, if we extend the model given by equation (5.3) to have two layers of feature mappings, the resulting model would become

$$f(\mathbf{x}) = \frac{1}{K} \boldsymbol{\alpha}^\top \mathbf{h}^{(2)} \left( \mathbf{h}^{(1)}(\mathbf{x}) \right). \quad (5.7)$$

---

<sup>1</sup>The above derivation gives the same result if the parameters of the hidden units are random. However, to avoid confusion, we refer to layers with infinitely-many nodes as “fixed”, since their distribution on outputs always the same with probability one.

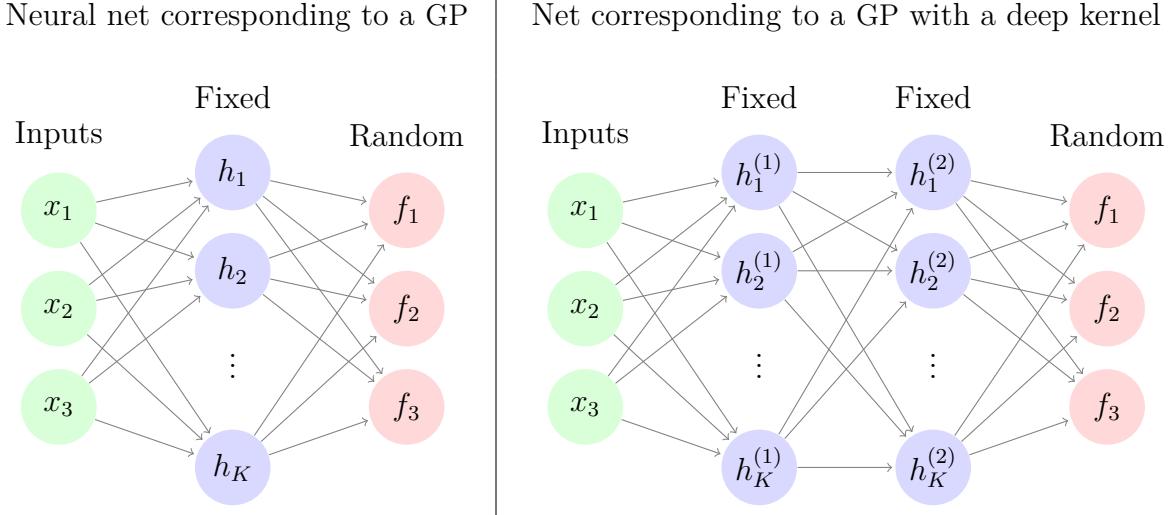


Figure 5.1 *Left:* GPs can be understood as a one-hidden-layer MLP with infinitely many fixed hidden units having unknown weights. *Right:* Multiple layers of fixed hidden units gives rise to a GP with a deep kernel, but not a deep GP.

If the features  $\mathbf{h}^{(1)}(\mathbf{x})$  and  $\mathbf{h}^{(2)}(\mathbf{x})$  are considered fixed with only the last-layer weights  $\boldsymbol{\alpha}$  unknown, this model corresponds to a (non-deep) GP with a “deep kernel”, given by

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{h}^{(2)}(\mathbf{h}^{(1)}(\mathbf{x})))^\top \mathbf{h}^{(2)}(\mathbf{h}^{(1)}(\mathbf{x}')) \quad (5.8)$$

GPs with deep kernels, explored in section 5.5, imply a fixed representation as opposed to a prior over representations. Thus, unless we richly parameterize these kernels, their capacity to learn an appropriate representation is limited in comparison to more flexible models such as deep neural networks, or deep GPs.

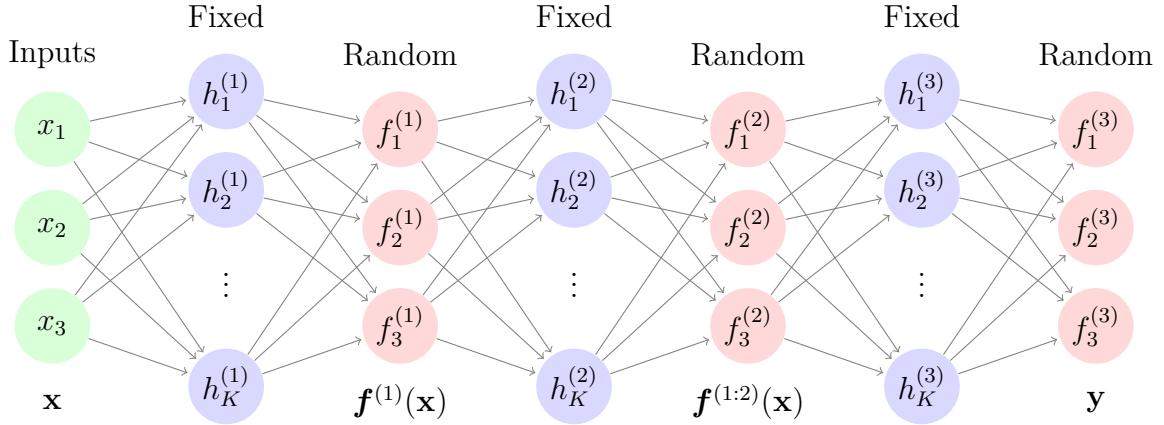
### 5.1.3 Two network architectures equivalent to deep GPs

There are two equivalent neural net architectures which correspond to deep GPs: one with fixed non-linearities, and another with GP-distributed nonlinearities.

To construct a neural network with fixed nonlinearities that corresponds to a deep GP, one can start with the infinitely-wide deep GP shown in figure 5.1(right), and introduce a finite set of nodes in between each infinitely-wide set of fixed basis functions. This architecture is shown in the top of figure 5.2. The  $D^{(\ell)}$  outputs  $\mathbf{f}^{(\ell)}(\mathbf{x})$  in between each layer are weighted sums (with random weights) of the fixed hidden units of the layer below, and the next layer’s hidden units depend only on these  $D^{(\ell)}$  outputs.

This alternating-layer architecture has an interpretation as a series of linear informa-

A neural net with fixed activation functions corresponding to a 3-layer deep GP



A net with nonparametric activation functions corresponding to a 3-layer deep GP

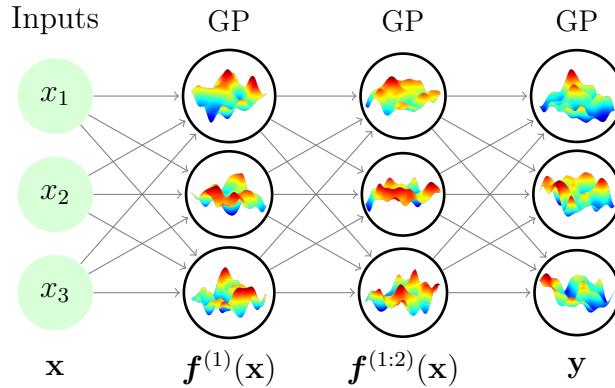


Figure 5.2 Two equivalent views of deep GPs as neural networks. *Top:* A neural network whose every second layer is a weighted sum of an infinite number of fixed hidden units, where the weights are initially unknown. *Bottom:* A neural network with a finite number of hidden units, each with a different unknown non-parametric activation function. The activation functions are visualized by draws from 2-dimensional GPs, although their input dimension will actually be the same as the output dimension of the previous layer.

tion bottlenecks. To see this, we can simply substitute equation (5.3) into equation (5.6) to get

$$\mathbf{h}^{(\ell)}(\mathbf{x}) = \sigma \left( \mathbf{b}^{(\ell)} + \mathbf{W}^{(\ell)} \mathbf{V}^{(\ell-1)} \mathbf{h}^{(\ell-1)}(\mathbf{x}) \right). \quad (5.9)$$

Thus, ignoring the intermediate outputs  $\mathbf{f}^{(\ell)}(\mathbf{x})$ , a deep GP is an infinitely-wide, deep MLP with each pair of layers connected by random, rank- $D_\ell$  matrices  $\mathbf{W}^{(\ell)} \mathbf{V}^{(\ell-1)}$ .

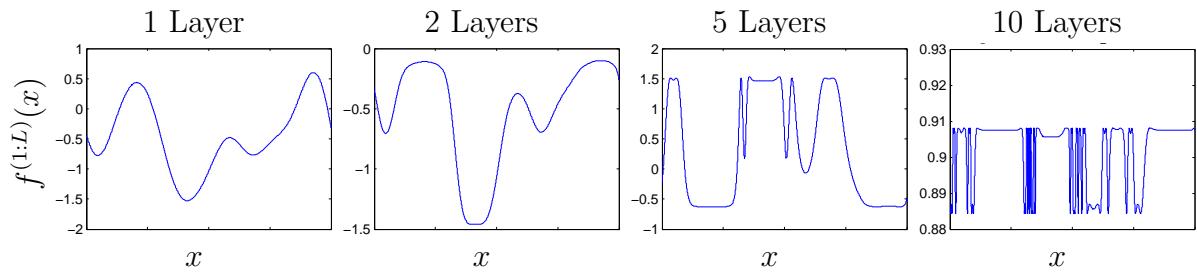


Figure 5.3 A one-dimensional draw from a deep GP prior, shown at different depths. The  $x$ -axis is the same for all plots. After a few layers, the functions begin to be either nearly flat, or highly varying, everywhere. This is a consequence of the distribution on derivatives becoming heavy-tailed. The function values also tend to cluster around the same few values as the depth increases. This happens because once the function values in a particular region are mapped to the same value in an intermediate layer, there is no way for them to be mapped to different values in later layers.

The second, more direct way to construct a network architecture corresponding to a deep GP is to integrate out all  $\mathbf{V}^{(\ell)}$ , and view deep GPs as a neural network with a finite number of nonparametric, GP-distributed basis functions at each layer, in which  $\mathbf{f}^{(1:\ell)}(\mathbf{x})$  represent the output of the hidden nodes at the  $\ell^{\text{th}}$  layer. This second view lets us compare deep GP models to standard neural net architectures more directly. Figure 5.1(bottom) shows this architecture.

## 5.2 Characterizing deep Gaussian process priors

In this section, we develop several theoretical results that explore the behavior of deep GPs as a function of their depth. This will allow us in section 5.3 to formally identify a pathology that emerges in very deep networks.

Specifically, we will show that the size of the derivative of a one-dimensional deep GP becomes log-normal distributed as the network becomes deeper. We will also show that the Jacobian of a multivariate deep GP is a product of independent Gaussian matrices with independent entries.

### 5.2.1 One-dimensional asymptotics

In this section, we derive the limiting distribution of the derivative of an arbitrarily deep, one-dimensional GP with a squared-exp kernel:

$$\text{SE}(x, x') = \sigma^2 \exp\left(\frac{-(x - x')^2}{2w^2}\right). \quad (5.10)$$

The parameter  $\sigma^2$  controls the variance of functions drawn from the prior, and the lengthscale parameter  $w$  controls the smoothness. The derivative of a GP with a squared-exp kernel is point-wise distributed as  $\mathcal{N}(0, \sigma^2/w^2)$ . Intuitively, a draw from a GP is likely to have large derivatives if the kernel has high variance and small lengthscales.

By the chain rule, the derivative of a one-dimensional deep GP is simply a product of the derivatives of each layer, which are drawn independently. The distribution of the absolute value of this derivative is a product of half-normals, each with mean  $\sqrt{2\sigma^2/\pi w^2}$ . If we choose kernel parameters so that  $\sigma^2/w^2 = \pi/2$ , then the expected magnitude of the derivative remains constant regardless of the depth.

The log of the magnitude of the derivatives has moments

$$\begin{aligned} m_{\log} &= \mathbb{E} \left[ \log \left| \frac{\partial f(x)}{\partial x} \right| \right] = 2 \log \left( \frac{\sigma}{w} \right) - \log 2 - \gamma \\ v_{\log} &= \mathbb{V} \left[ \log \left| \frac{\partial f(x)}{\partial x} \right| \right] = \frac{\pi^2}{4} + \frac{\log^2 2}{2} - \gamma^2 - \gamma \log 4 + 2 \log \left( \frac{\sigma}{w} \right) \left[ \gamma + \log 2 - \log \left( \frac{\sigma}{w} \right) \right] \end{aligned} \quad (5.11)$$

where  $\gamma \approx 0.5772$  is Euler's constant. Since the second moment is finite, by the central limit theorem, the limiting distribution of the size of the gradient approaches log-normal as  $L$  grows:

$$\log \left| \frac{\partial f^{(1:L)}(x)}{\partial x} \right| = \sum_{\ell=1}^L \log \left| \frac{\partial f^{(\ell)}(x)}{\partial x} \right| \xrightarrow{L \rightarrow \infty} \mathcal{N}(Lm_{\log}, L^2v_{\log}) \quad (5.12)$$

Even if the expected magnitude of the derivative remains constant, the variance of the log-normal distribution grows without bound as the depth increases.

Because the log-normal distribution is heavy-tailed and its domain is bounded below by zero, the derivative will become very small almost everywhere, with rare but very large jumps. Figure 5.3 shows this behavior in a draw from a 1D deep GP prior, at varying depths. This figure also shows that once the derivative in one region of the input space becomes very large or very small, it is likely to remain that way in subsequent layers.

### 5.2.2 Distribution of the Jacobian

Next, we characterize the distribution on Jacobians of multivariate functions drawn from a deep GP prior.

**Lemma 5.2.1.** *The partial derivatives of a function mapping  $\mathbb{R}^D \rightarrow \mathbb{R}$  drawn from a GP prior with a product kernel are independently Gaussian distributed.*

*Proof.* Because differentiation is a linear operator, the derivatives of a function drawn from a GP prior are also jointly Gaussian distributed. The covariance between partial derivatives with respect to input dimensions  $d_1$  and  $d_2$  of vector  $\mathbf{x}$  are given by Solak et al. (2003):

$$\text{cov}\left(\frac{\partial f(\mathbf{x})}{\partial x_{d_1}}, \frac{\partial f(\mathbf{x})}{\partial x_{d_2}}\right) = \frac{\partial^2 k(\mathbf{x}, \mathbf{x}')}{\partial x_{d_1} \partial x'_{d_2}} \Big|_{\mathbf{x}=\mathbf{x}'} \quad (5.13)$$

If our kernel is a product over individual dimensions  $k(\mathbf{x}, \mathbf{x}') = \prod_d^D k_d(x_d, x'_d)$ , then the off-diagonal entries are zero, implying that all elements are independent.  $\square$

For example, in the case of the multivariate squared-exp kernel, the covariance between derivatives has the form:

$$\begin{aligned} f(\mathbf{x}) &\sim \text{GP}\left(0, \sigma^2 \prod_{d=1}^D \exp\left(-\frac{1}{2} \frac{(x_d - x'_d)^2}{w_d^2}\right)\right) \\ \implies \text{cov}\left(\frac{\partial f(\mathbf{x})}{\partial x_{d_1}}, \frac{\partial f(\mathbf{x})}{\partial x_{d_2}}\right) &= \begin{cases} \frac{\sigma^2}{w_{d_1}^2} & \text{if } d_1 = d_2 \\ 0 & \text{if } d_1 \neq d_2 \end{cases} \end{aligned} \quad (5.14)$$

**Lemma 5.2.2.** *The Jacobian of a set of  $D$  functions  $\mathbb{R}^D \rightarrow \mathbb{R}$  drawn from independent GP priors, each having product kernel is a  $D \times D$  matrix of independent Gaussian R.V.'s*

*Proof.* The Jacobian of the vector-valued function  $\mathbf{f}(\mathbf{x})$  is a matrix  $J$  with elements  $J_{ij}(\mathbf{x}) = \frac{\partial f_i(\mathbf{x})}{\partial x_j}$ . Because the GPs on each output dimension  $f_d(\mathbf{x})$  are independent, it follows that each row of  $J$  is independent. Lemma 5.2.1 shows that the elements of each row are independent Gaussian. Thus all entries in the Jacobian of a GP-distributed transform are independent Gaussian R.V.'s.  $\square$

**Theorem 5.2.3.** *The Jacobian of a deep GP with a product kernel is a product of independent Gaussian matrices, with each entry in each matrix being drawn independently.*

*Proof.* When composing  $L$  different functions, we denote the *immediate* Jacobian of the function mapping from layer  $\ell - 1$  to layer  $\ell$  as  $J^{(\ell)}(\mathbf{x})$ , and the Jacobian of the entire composition of  $L$  functions by  $J^{(1:L)}(\mathbf{x})$ . By the multivariate chain rule, the Jacobian of a composition of functions is simply the product of the immediate Jacobian matrices of each function. Thus the Jacobian of the composed (deep) function  $\mathbf{f}^{(L)}(\mathbf{f}^{(L-1)}(\dots \mathbf{f}^{(3)}(\mathbf{f}^{(2)}(\mathbf{f}^{(1)}(\mathbf{x})))\dots))$  is

$$J^{(1:L)}(\mathbf{x}) = J^{(L)} J^{(L-1)} \dots J^{(3)} J^{(2)} J^{(1)}. \quad (5.15)$$

By lemma 5.2.2, each  $J_{i,j}^{(\ell)} \stackrel{\text{ind}}{\sim} \mathcal{N}$ , so the complete Jacobian is a product of independent Gaussian matrices, with each entry of each matrix drawn independently.  $\square$

Theorem 5.2.3 allows us to analyze the representational properties of a deep Gaussian process by examining the properties of products of independent Gaussian matrices.

### 5.3 Formalizing a pathology

Rifai et al. (2011a) argue that good representations of data manifolds are invariant in directions orthogonal to the manifold. Conversely, a good representation must also change in directions tangent to the data manifold, in order to preserve relevant information. Figure 5.4 visualizes this idea.

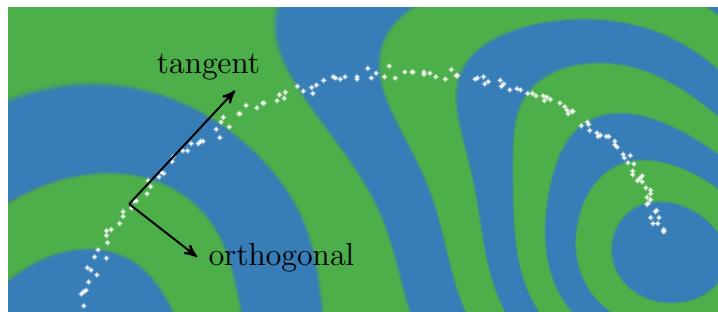


Figure 5.4 Representing a 1-D data manifold. Colors correspond to the computed representation as a function of the input space. The representation (blue & green) varies in directions tangent to the data manifold (white), preserving information for later layers. The representation changes little in directions orthogonal to the manifold, making it robust to noise in those directions.

As in Rifai et al. (2011b), we characterize the representational properties of a function by the singular value spectrum of the Jacobian. The number of relatively large singular

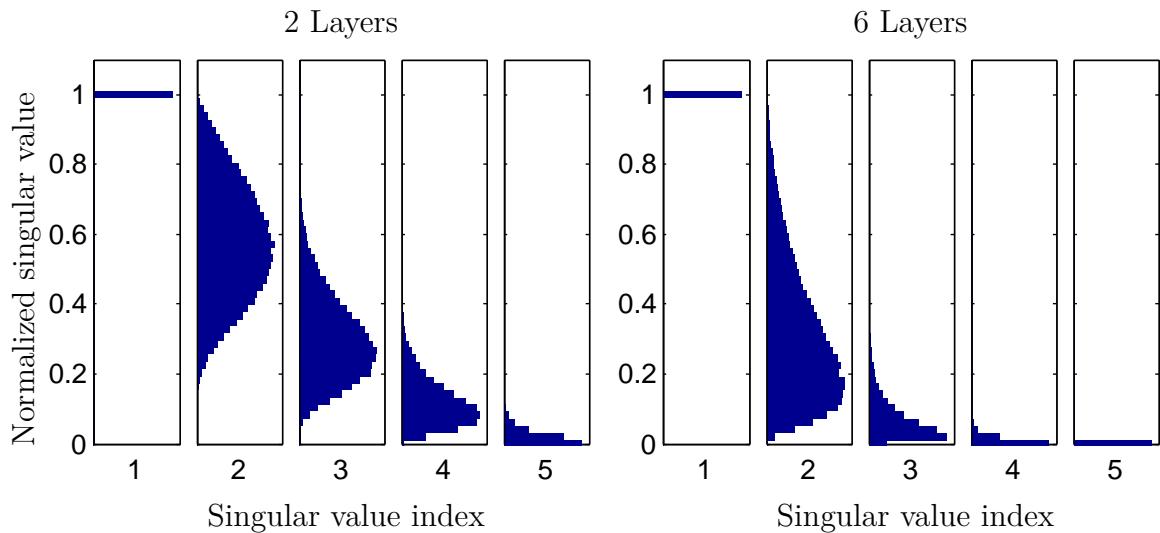


Figure 5.5 The distribution of normalized singular values of the Jacobian of a function drawn from a 5-dimensional deep GP prior 25 layers deep (*Left*) and 50 layers deep (*Right*). As nets get deeper, the largest singular value tends to become much larger than the others. This implies that with high probability, these functions vary little in all directions but one, making them unsuitable for computing representations of manifolds of more than one dimension.

values of the Jacobian roughly correspond to the number of directions in which the representation varies. Rifai et al. (2011b) analyzed the Jacobian at location of the training points, but because the priors we are examining are stationary, the distribution of the Jacobian is identical everywhere. Figure 5.5 shows the singular value spectrum for 5-dimensional deep GPs of different depths. As the net gets deeper, the largest singular value dominates, implying there is usually only one effective degree of freedom in representation being computed.

Figure 5.6 demonstrates a related pathology that arises when composing functions to produce a deep density model. The density in the observed space eventually becomes locally concentrated onto one-dimensional manifolds, or *filaments*, again implying that such models are unsuitable to model manifolds whose underlying dimensionality is greater than one.

To visualize this pathology in another way, figure 5.7 illustrates a color-coding of the representation computed by a deep GP, evaluated at each point in the input space. After 10 layers, we can see that locally, there is usually only one direction that one can move in  $\mathbf{x}$ -space in order to change the value of the computed representation. This means that such representations are likely to be unsuitable for decision tasks that depend on

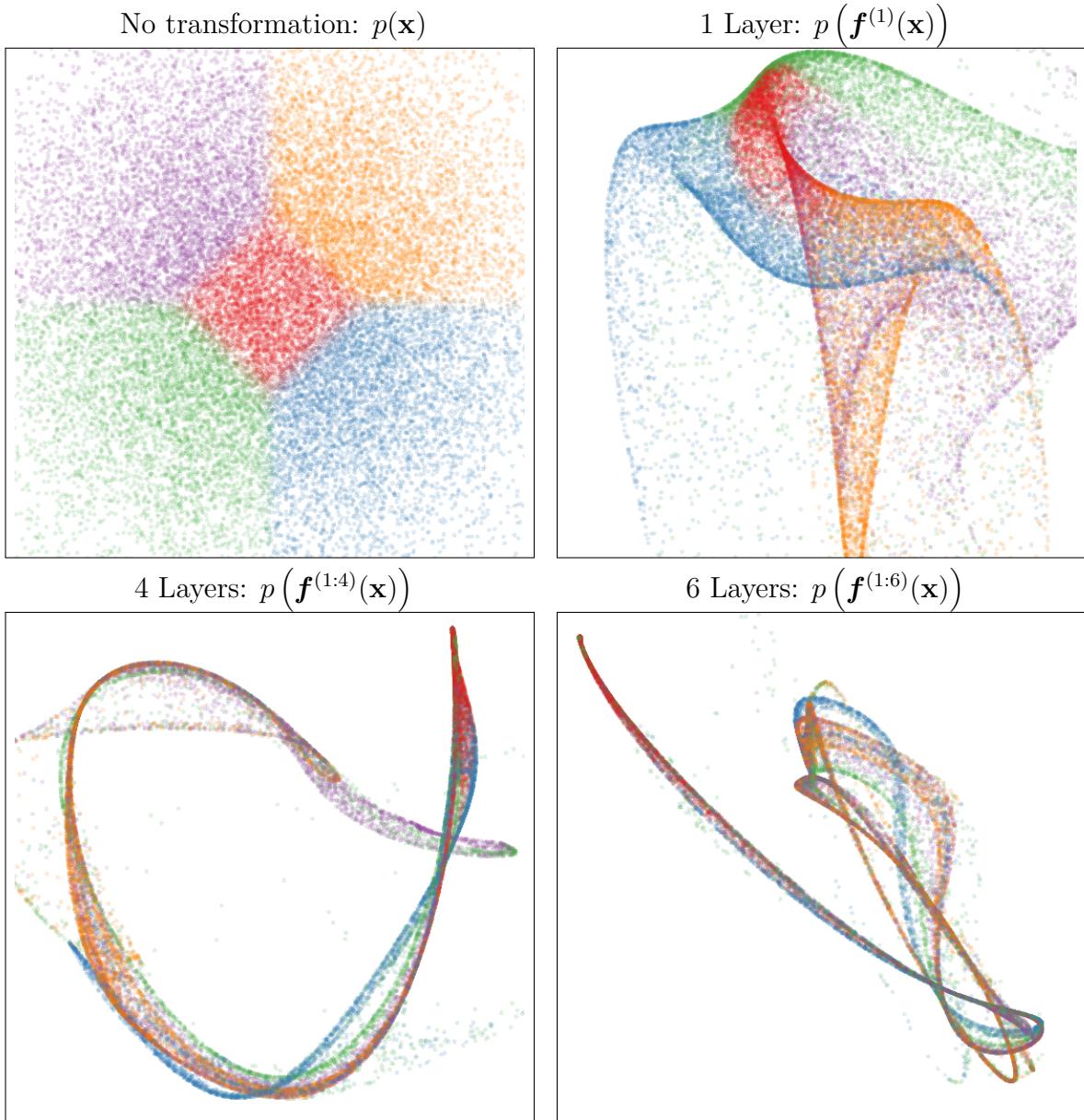


Figure 5.6 Points warped by a function drawn from a deep GP prior. *Top left:* Points drawn from a 2-dimensional Gaussian distribution, color-coded by their location. *Subsequent panels:* Those same points, successively warped by functions drawn from a GP prior. As the number of layers increases, the density concentrates along one-dimensional filaments.

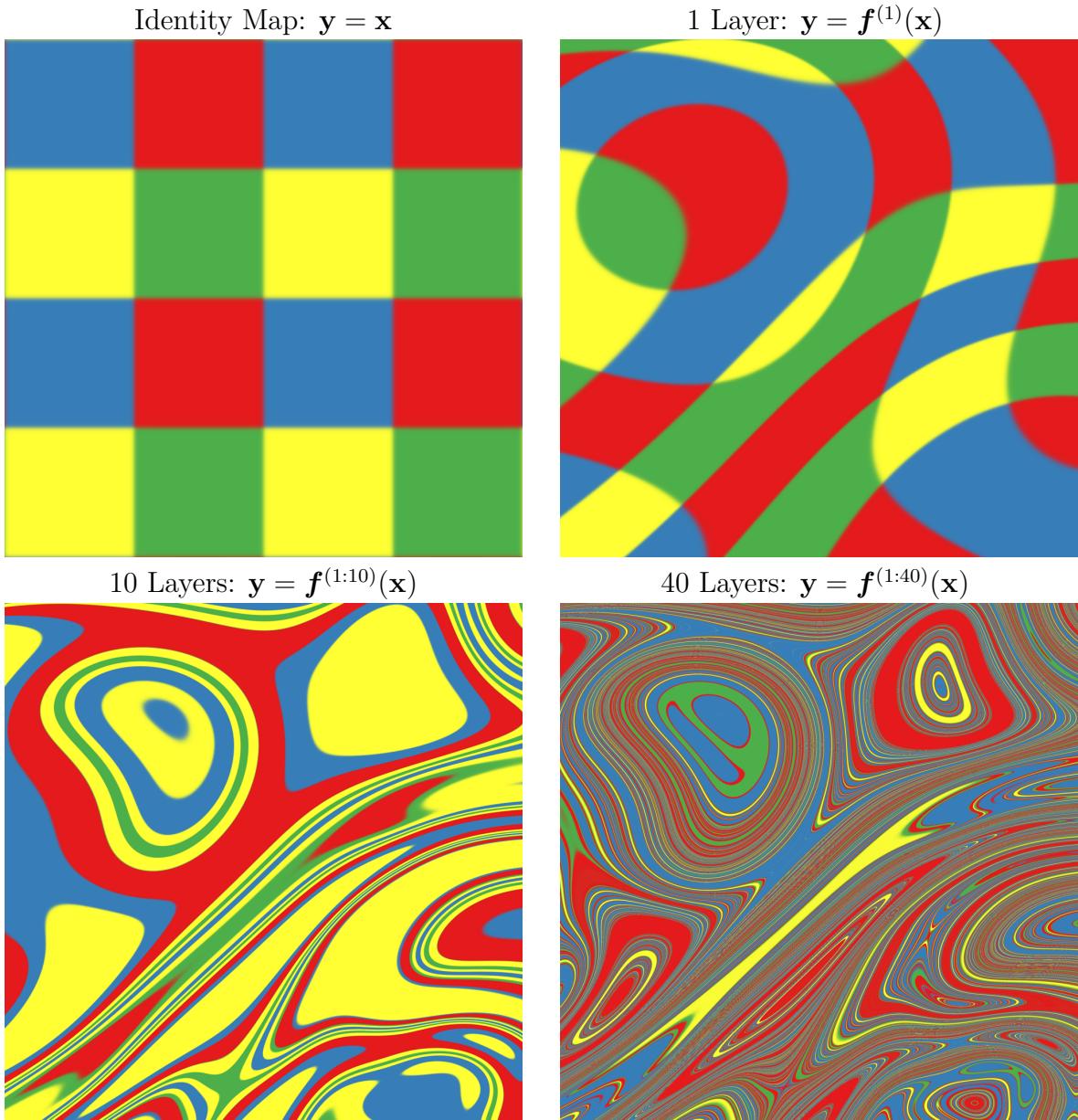


Figure 5.7 A visualization of the feature map implied by a draw from a deep GP. Colors correspond to the location  $\mathbf{y} = \mathbf{f}(\mathbf{x})$  that each point is mapped to after being warped by a deep GP. The number of directions in which the color changes rapidly corresponds to the number of large singular values in the Jacobian. Just as the densities in figure 5.6 became locally one-dimensional, there is usually only one direction that one can move  $\mathbf{x}$  in locally to change  $\mathbf{y}$ . This means that  $\mathbf{f}$  is unlikely to be a suitable representation for decision tasks that depend on more than one aspect of  $\mathbf{x}$ . Also note that the overall shape of the mapping remains the same as the number of layers increase. For example, the roughly circular shape remains in the top-left corner even after 40 independent warpings.

more than one property of the input.

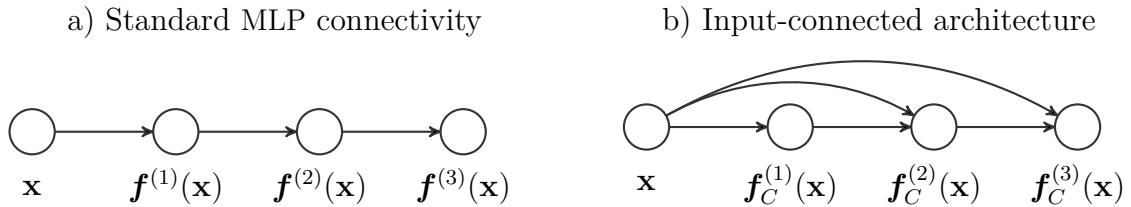
To what extent are these pathologies present in the types of neural networks commonly used in practice? In simulations, we found that for deep functions with a fixed latent dimension  $D$ , the singular value spectrum remained relatively flat for hundreds of layers as long as  $D > 100$ . Thus, these pathologies are unlikely to severely effect the relatively shallow, wide networks most commonly used in practice.

## 5.4 Fixing the pathology

As suggested by [Neal \(1995, chapter 2\)](#), we can fix the pathologies exhibited in figures [figure 5.6](#) and [5.7](#) by simply making each layer depend not only on the output of the previous layer, but also on the original input  $\mathbf{x}$ . We refer to these models as *input-connected* networks, and denote deep functions having this architecture with the subscript  $C$ , as in  $f_C(\mathbf{x})$ . Figure [5.8](#) shows a graphical representation of the two connectivity architectures. Similar connections between non-adjacent layers can also be found the primate visual cortex ([Maunsell and van Essen, 1983](#)). Formally, this functional dependence can be written as

$$f_C^{(1:L)}(\mathbf{x}) = \mathbf{f}^{(L)} \left( f_C^{(1:L-1)}(\mathbf{x}), \mathbf{x} \right), \quad \forall L \quad (5.16)$$

Visualizations of the resulting prior on functions are shown in [figures 5.9](#), [5.10](#) and [5.12](#).



[Figure 5.8](#) Two different architectures for deep neural networks. *Left:* The standard architecture connects each layer’s outputs to the next layer’s inputs. *Right:* The input-connected architecture also connects the original input  $\mathbf{x}$  to each layer.

The Jacobian of an input-connected deep function is defined by the recurrence

$$J_C^{(1:L)} = J^{(L)} \begin{bmatrix} J_C^{(1:L-1)} \\ I_D \end{bmatrix}. \quad (5.17)$$

which, in the case of a deep GP, is still a product of independent Gaussian matrices.

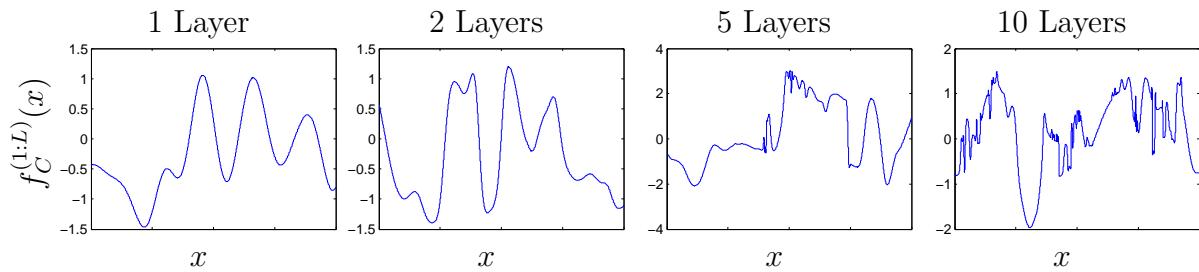


Figure 5.9 A draw from a 1D deep GP prior with each layer connected to the input. The  $x$ -axis is the same for all plots. Even after many layers, the functions remain smooth in some regions, while varying rapidly in other regions. Compare to standard-connectivity deep GP draws shown in figure 5.3.

Figure 5.11 shows that with this architecture, even 50-layer deep GPs have well-behaved singular value spectra.

The pathology examined in this section is an example of the sort of analysis made possible by a well-defined prior on functions. As explained in section 5.1, GPs are examples of a particular limit of Bayesian neural networks. The figures and analysis done in this section could have also been done using Bayesian neural networks with finite numbers of nodes, but would have been more difficult. In particular, care would need to be taken to ensure that the networks do not produce degenerate mappings due to saturation of the hidden units.

## 5.5 Deep kernels

Bengio et al. (2006) showed that kernel machines have limited generalization ability when they use “local” kernels, such as the squared-exp. However, as we saw in chapters 2 and 3, structured kernels can be constructed through addition and multiplication which allow non-local extrapolation.

We can also build non-local kernels by composing fixed feature maps. To return to an example given in section 2.7, periodic kernels can be viewed as a 2-layer-deep kernel, in which the first layer maps  $x \rightarrow [\sin(x), \cos(x)]$ , and the second layer maps through basis functions corresponding to the SE kernel.

We can construct other useful kernels by composing fixed feature maps several times, creating deep kernels. This section builds on the work of Cho and Saul (2009), who derived several kinds of deep kernels by applying multiple layers of feature mappings.

In principle, we can compose the feature mapping of any two kernels to get a new

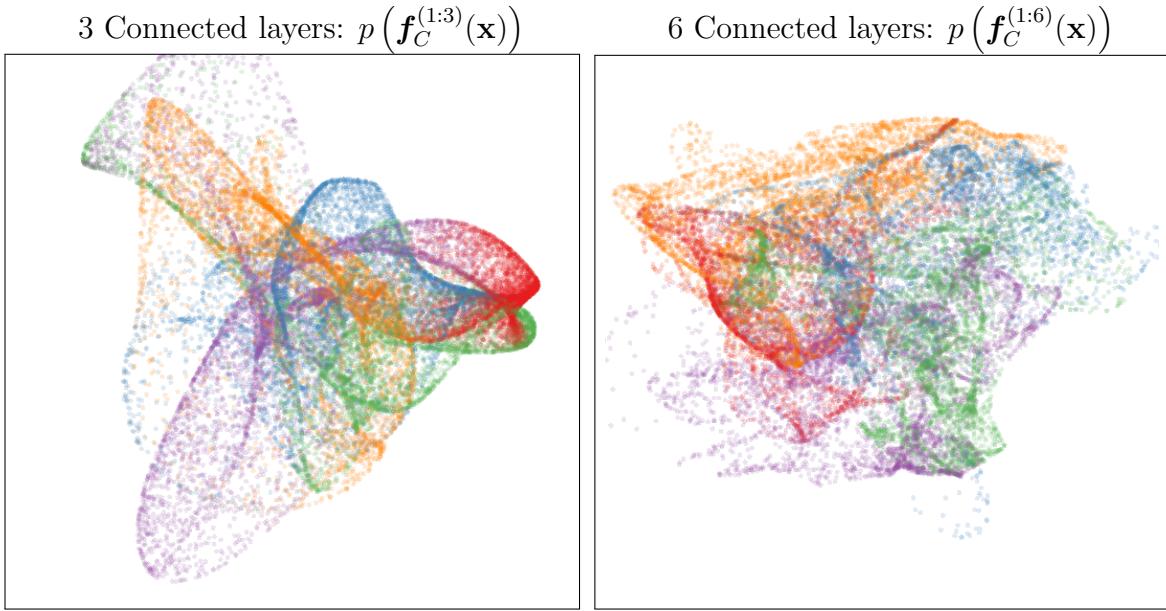


Figure 5.10 Points warped by a draw from a deep GP, with each layer connected to the input  $\mathbf{x}$ . As depth increases, the density becomes more complex without concentrating everywhere along one-dimensional filaments.

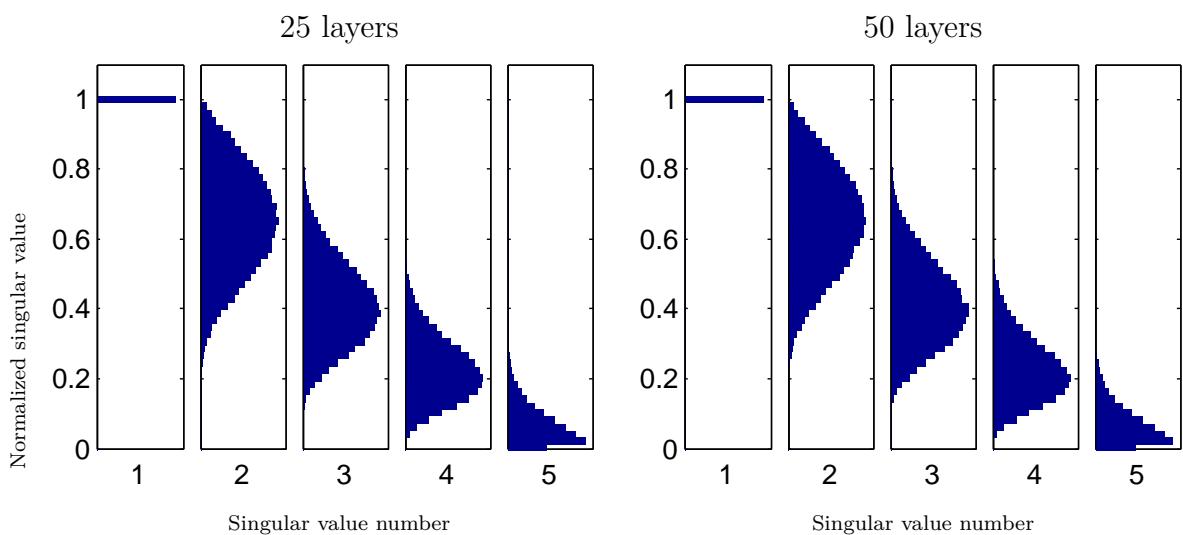


Figure 5.11 The distribution of singular values drawn from 5-dimensional input-connected deep GP priors, 25 and 50 layers deep. The singular values remain roughly the same scale as one another, meaning that the model outputs are sensitive to most directions of variation in the input.

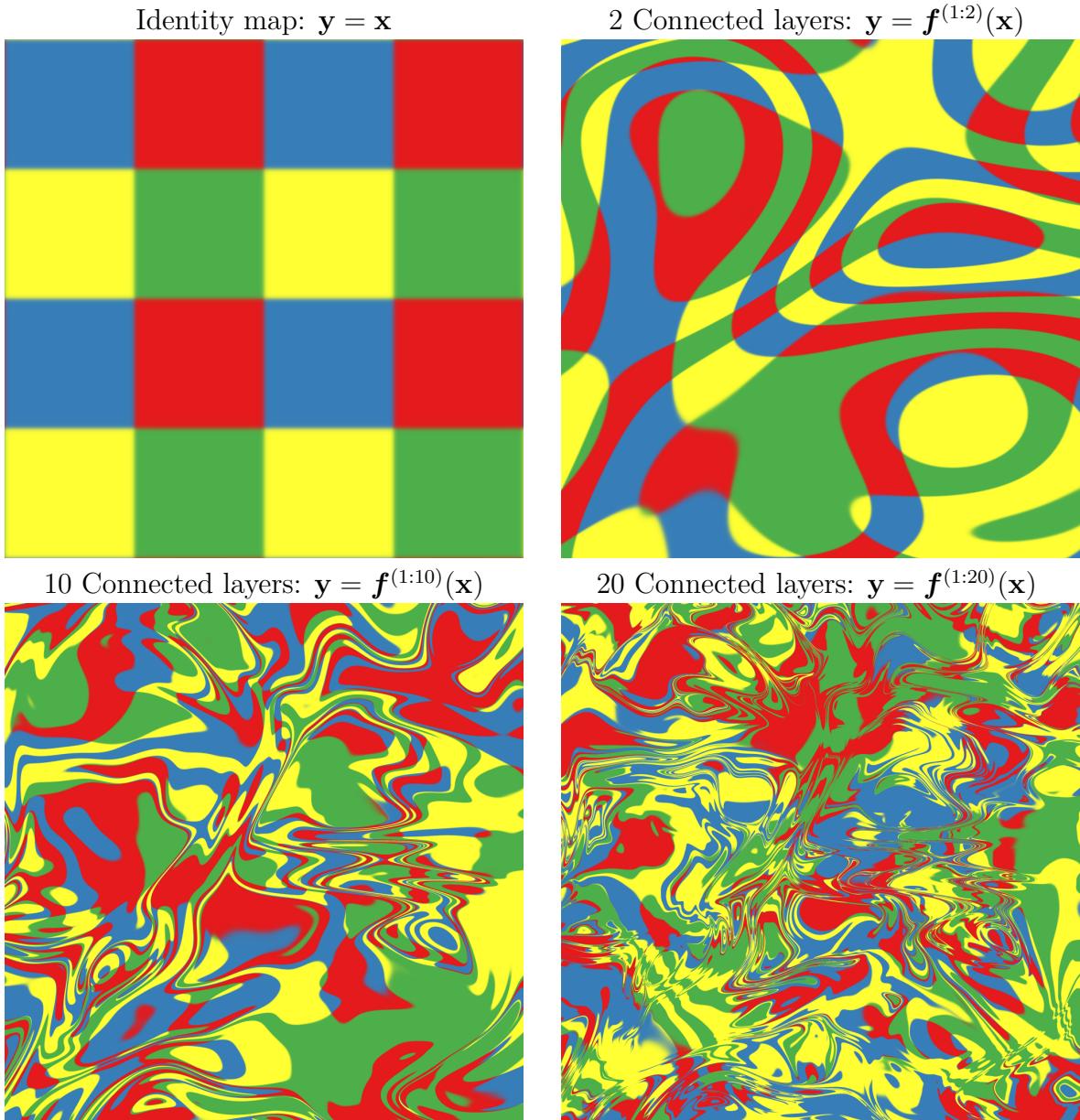


Figure 5.12 Feature mapping of a deep GP with each layer connected to the input  $\mathbf{x}$ . Compare to the mappings shown in figure 5.7. Just as the densities in figure 5.10 remained locally two-dimensional even after many warpings, in the mapping shown here there are sometimes two directions that one can move locally in  $\mathbf{x}$  to in order to change the values of  $\mathbf{f}(\mathbf{x})$ . This means that the input-connected prior puts mass on a greater variety of types of representations, some of which depend on all aspects of the input.

one:

$$k_a(\mathbf{x}, \mathbf{x}') = \mathbf{h}_a(\mathbf{x})^\top \mathbf{h}_a(\mathbf{x}') \quad (5.18)$$

$$k_b(\mathbf{x}, \mathbf{x}') = \mathbf{h}_b(\mathbf{x})^\top \mathbf{h}_b(\mathbf{x}') \quad (5.19)$$

$$(k_b \circ k_a)(\mathbf{x}, \mathbf{x}') = k_b(\mathbf{h}_a(\mathbf{x}), \mathbf{h}_a(\mathbf{x}')) = [\mathbf{h}_b(\mathbf{h}_a(\mathbf{x}))]^\top \mathbf{h}_b(\mathbf{h}_a(\mathbf{x}')) \quad (5.20)$$

However, this composition might not always have a closed form if the number of hidden features of either kernel in infinite.

Fortunately, composing the squared-exp kernel with any implicit mapping given by another kernel has a simple closed form. If  $k(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}')$ , then

$$\begin{aligned} (\text{SE} \circ k)(\mathbf{x}, \mathbf{x}') &= k_{\text{SE}}(\mathbf{h}(\mathbf{x}), \mathbf{h}(\mathbf{x}')) = \\ &= \exp\left(-\frac{1}{2}\|\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x}')\|_2^2\right) \\ &= \exp\left(-\frac{1}{2}\left[\mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}) - 2\mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}') + \mathbf{h}(\mathbf{x}')^\top \mathbf{h}(\mathbf{x}')\right]\right) \\ &= \exp\left(-\frac{1}{2}\left[k(\mathbf{x}, \mathbf{x}) - 2k(\mathbf{x}, \mathbf{x}') + k(\mathbf{x}', \mathbf{x}')\right]\right). \end{aligned} \quad (5.21)$$

This formula lets expresses the composed kernel function  $\text{SE} \circ k$  exactly in terms of evaluations of the original kernel function  $k$ .

### 5.5.1 Infinitely deep kernels

What happens when we repeat this composition of feature maps many times, starting with the squared-exp kernel? If the output variance  $k(\mathbf{x}, \mathbf{x}) = 1$ , then in the infinite limit, this operation converges to  $(\text{SE} \circ \text{SE} \circ \text{SE} \circ \dots \text{SE})(\mathbf{x}, \mathbf{x}') = 1$  for all pairs of inputs, which corresponds to a prior on constant functions  $f(\mathbf{x}) = c$ .

As before, we can overcome this degeneracy by connecting the inputs  $\mathbf{x}$  to each layer. To do so, we simply concatenate the feature vector at each layer,  $\mathbf{h}^{(\ell)}(\mathbf{x})$ , with the input vector  $\mathbf{x}$ :

$$\begin{aligned} k^{(\ell+1)}(\mathbf{x}, \mathbf{x}') &= \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} \mathbf{h}^{(\ell)}(\mathbf{x}) \\ \mathbf{x} \end{bmatrix} - \begin{bmatrix} \mathbf{h}^{(\ell)}(\mathbf{x}') \\ \mathbf{x}' \end{bmatrix}\right\|_2^2\right) \\ &= \exp\left(-\frac{1}{2}\left[k^{(\ell)}(\mathbf{x}, \mathbf{x}) - 2k^{(\ell)}(\mathbf{x}, \mathbf{x}') + k^{(\ell)}(\mathbf{x}', \mathbf{x}') - \|\mathbf{x} - \mathbf{x}'\|_2^2\right]\right) \quad (5.22) \end{aligned}$$

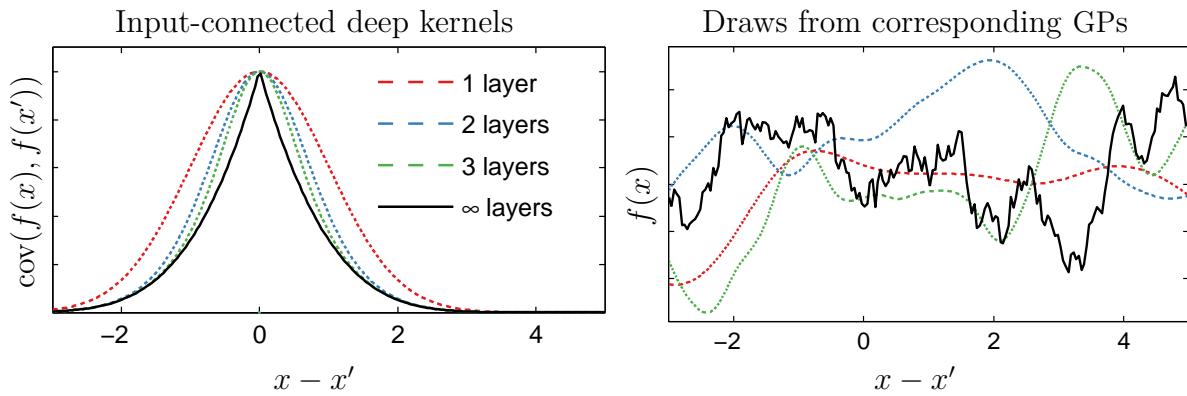


Figure 5.13 *Left*: Input-connected deep kernels of different depths. By connecting the inputs  $\mathbf{x}$  to each layer, the kernel can still depend on its input even after arbitrarily many layers of computation. *Right*: Draws from GPs with deep input-connected kernels.

Starting with the squared-exp kernel, this repeated mapping satisfies

$$k^{(\infty)}(\mathbf{x}, \mathbf{x}') - \log(k^{(\infty)}(\mathbf{x}, \mathbf{x}')) = 1 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2. \quad (5.23)$$

The solution to this recurrence has no closed form, but has a similar shape to the Ornstein-Uhlenbeck covariance  $\text{OU}(x, x') = \exp(-|x - x'|)$  but with lighter tails. Samples from a GP prior with this kernel are not differentiable, and are locally fractal. Figure 5.13 shows this kernel at different depths, as well as samples from the resulting GP prior.

We can also consider two other connectivity architectures: one in which each layer is connected to the output layer, and another in which every layer is connected to all subsequent layers. It is easy to show that in the limit of infinite depth, both these architectures converge to  $k(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x}, \mathbf{x}')$ , a white noise kernel.

### 5.5.2 When are deep kernels useful models?

Kernels correspond to fixed feature maps, and so kernel learning is an example of implicit representation learning. As we saw in chapters 2 and 3, kernels can capture rich structure and can enable many types of generalization. The relatively uninteresting properties of the deep kernels derived in this section simply reflect the fact that an arbitrary computation, even if it is “deep”, is not likely to give rise to a useful representation, unless combined with learning. To put it another way, any fixed representation is unlikely to be useful unless it has been chosen specifically for the problem at hand.

## 5.6 Related work

### Deep Gaussian processss

Neal (1995, chapter 2) discussed the properties of arbitrarily deep Bayesian neural networks, including those that would give rise to deep GPs. He noted that infinitely deep random neural networks without extra connections to the input would be equivalent to a Markov chain, and therefore would lead to degenerate priors on functions. He also suggested connecting the input to each layer in order to fix this problem. Much of the analysis in this chapter can be seen as a detailed investigation and vindication of these claims.

The first instance of deep GPs being used in practice was (Lawrence and Moore, 2007), who presented a model called “hierarchical GP-LVMs”, in which time was mapped through a composition of multiple GPs to produce observations.

The term “deep Gaussian processes” was first used by Damianou and Lawrence (2013), who also developed a variational inference method, analyzed the effect of automatic relevance determination, and showed that deep GPS could learn with relatively little data. They used the term “deep GP” to refer both to supervised models (compositions of GPs) and to unsupervised models (compositions of GP-LVMs). This conflation may be reasonable, since the activations of the hidden layers are themselves latent variables. Depending on kernel parameters, these latent variables may or may not depend on the inputs in the layer below. In general, supervised models can also be latent-variable models. For example, Wang and Neal (2012) investigated single-layer GP regression models that had additional latent inputs.

### Nonparametric neural networks

Adams et al. (2010) proposed a prior on arbitrarily deep Bayesian networks with an unknown and unbounded number of hidden units in each layer. Their architecture has connections only between adjacent layers, and may also be expected to have similar pathologies to those of deep GPs as the number of layers increases.

Deep density networks (Rippel and Adams, 2013) are constructed through a series of warpings of fixed output dimension, with penalty terms encouraging the preservation of information about lower layers. This is another promising approach to fixing the pathology discussed in section 5.3.

Wilson et al. (2012) introduced Gaussian process regression networks, which are defined as a matrix product of draws from GPs priors, rather than a composition. These

networks have the form:

$$\mathbf{y}(\mathbf{x}) = \mathbf{W}(\mathbf{x})\mathbf{f}(\mathbf{x}) \quad \text{with each } f_d, W_{d,j} \stackrel{\text{iid}}{\sim} \mathcal{GP}(\mathbf{0}, \text{SE} + \text{WN}). \quad (5.24)$$

We can easily define a “deep” Gaussian process regression network:

$$\mathbf{y}(\mathbf{x}) = \mathbf{W}^{(3)}(\mathbf{x})\mathbf{W}^{(2)}(\mathbf{x})\mathbf{W}^{(1)}(\mathbf{x})\mathbf{f}(\mathbf{x}) \quad (5.25)$$

which repeatedly adds and multiplies functions drawn from GPs, in contrast to deep GPs, which repeatedly compose functions. This prior on functions has a similar form to the Jacobian of a deep GP (equation (5.15)), and so might be amenable to a similar analysis to that of section 5.2.

## Recurrent networks

Bengio et al. (1994) and Pascanu et al. (2012) analyze a related problem with gradient-based learning in recurrent nets, the “exploding-gradients” problem. They note that in recurrent neural networks, the size of the training gradient can grow or shrink exponentially as it is back-propagated, making gradient-based training difficult.

## Deep kernels

The first systematic examination of deep kernels was done by Cho and Saul (2009), who derived closed-form composition rules for SE, polynomial, and arc-cosine kernels, and showed that deep arc-cosine kernels performed competitively in machine-vision applications when used in a SVM.

Hermans and Schrauwen (2012) constructed deep kernels in a time-series setting, constructing kernels corresponding to infinite-width *recurrent* neural networks. They also proposed concatenating the implicit feature vectors from previous time-steps with the current inputs, resulting in an architecture analogous to the input-connected architecture proposed by Neal (1995, chapter 2).

## Analyses of deep learning

Montavon et al. (2010) performed a layer-wise analysis of deep networks, and noted that the performance of MLPs degrades as the number of layers with random weights increases.

The experiments of Saxe et al. (2011) suggested that most of the performance of convolutional neural networks could be attributed to the architecture alone. Later, Saxe et al. (2013) looked at the dynamics of gradient-based training methods in deep *linear* networks, as a tractable approximation to standard deep (nonlinear) neural networks.

### Source code

Source code to produce all figures is available at [github.com/duvenaud/deep-limits](https://github.com/duvenaud/deep-limits). The source code is also capable of producing visualizations of mappings such as figures 5.7 and 5.12 using neural nets instead of GPs at each layer.

## 5.7 Conclusions

This chapter showed that well-defined priors allow explicit examination of the assumptions being made about functions being modeled. As an example of the sort of analysis made possible this way, we attempted to gain insight into the properties of deep neural networks by characterizing the sorts of functions likely to be obtained under different choices of priors on compositions of functions.

First, we identified deep Gaussian processes as an easy-to-analyze model corresponding to multi-layer preceptrons with nonparametric activation functions. We then showed that representations based on repeated composition of independent functions exhibit a pathology where the representations becomes invariant to all directions of variation but one. Finally, we showed that this problem could be alleviated by connecting the input to each layer. We also examined properties of deep kernels, corresponding to arbitrarily many compositions of fixed features.

Much recent work on deep networks has focused on weight initialization (Martens, 2010), regularization (Lee et al., 2007) and network architecture (Gens and Domingos, 2013). However, the interactions between these different design decisions can be complex and difficult to characterize. If we can identify classes of priors that give our models desirable properties, these in turn may suggest regularization, initialization, and architecture choices that also provide such properties.

Existing neural network practice also requires expensive tuning of model hyperparameters such as the number of layers, the size of each layer, and regularization penalties by cross-validation. One advantage of deep GPs is that the approximate marginal likelihood allows a principled method for automatically determining such model details.

# Chapter 6

## Additive Gaussian Processes

In chapter 3, we showed how to learn the structure of a kernel by building it up piece-by-piece. This chapter presents an alternative approach, where we start with many different types of structure in the kernel, and adjust kernel parameters to discard whatever structures are *not* present in the current dataset. The advantage of this approach is that we do not need to run an expensive discrete-and-continuous optimization problem in order to build a structured model. Implementation is also much simpler.

The type of structure our kernel will represent are sums of functions of all possible combinations of input variables. We call this model class *additive Gaussian processes*. This model can be specified by a kernel which is a sum of all possible products of one-dimensional kernels.

There are  $2^D$  combinations of  $D$  inputs, so a naïve computation of such a kernel would be intractable. Furthermore, if each term has different kernel parameters, fitting or integrating over so many parameters would pose severe difficulty. To get around this problem, we introduce a parameterization of the kernel which allows efficient evaluation of all interaction terms. Empirically, this kernel has good predictive power in regression tasks, and its parameters are relatively interpretable.

The work in this chapter was done in collaboration with Hannes Nickisch and Carl Rasmussen, who derived and coded up the initial model. My role in the project was to examine the properties of the resulting model, clarify the connections to existing methods, to create all figures and run all experiments. That work was published in [Duvenaud et al. \(2011\)](#). The connection to dropout regularization is an independent original contribution.

## 6.1 Different types of multivariate additive structure

In chapter 2, we saw how additive structure in a GP prior enabled long-range extrapolation in multivariate regression problems. In general, models of the form

$$f(\mathbf{x}) = g(f(x_1) + f(x_2) + \cdots + f(x_D)) \quad (6.1)$$

are widely used in machine learning and statistics, partly because they are relatively easy to fit and interpret. Examples include logistic regression, linear regression, generalized linear models (Nelder and Wedderburn, 1972) and generalized additive models (Hastie and Tibshirani, 1990).

At the other end of the spectrum are models which allow the response to depend on all input variables simultaneously, having the most general form:

$$f(\mathbf{x}) = f(x_1, x_2, \dots, x_D) \quad (6.2)$$

An example would be a GP with an SE-ARD kernel. Such models are much more flexible than those having the form (6.1), but this flexibility can make it difficult to generalize to new combinations of input variables.

In between these extremes, we can consider function classes depending on pairs or triplets of inputs, such as

$$f(x_1, x_2, x_3) = f_{12}(x_1, x_2) + f_{23}(x_2, x_3) + f_{13}(x_1, x_3) \quad (6.3)$$

We call the number of input variables appearing in each term the *order* of a model class. Models of intermediate order such as (6.3) allow more flexibility than models of form (6.1) ( $D$ -th order), but have more structure than those of form (6.2) (first-order).

If the function being learned depends in some way on an interaction between all input variables, a  $D$ th-order term is required in order for the model to be consistent. However, if the function also contains lower-order interactions, capturing that structure will still improve the predictive performance on finite datasets.

## 6.2 Defining additive kernels

We now give a precise definition of the additive kernels introduced in this chapter. We first assign each dimension  $i \in \{1 \dots D\}$  a one-dimensional *base kernel*  $k_i(x_i, x'_i)$ . We then define the first order, second order and  $n$ th order additive kernel as:

$$k_{add_1}(\mathbf{x}, \mathbf{x}') = \sigma_1^2 \sum_{i=1}^D k_i(x_i, x'_i) \quad (6.4)$$

$$k_{add_2}(\mathbf{x}, \mathbf{x}') = \sigma_2^2 \sum_{i=1}^D \sum_{j=i+1}^D k_i(x_i, x'_i) k_j(x_j, x'_j) \quad (6.5)$$

$$k_{add_n}(\mathbf{x}, \mathbf{x}') = \sigma_n^2 \sum_{1 \leq i_1 < i_2 < \dots < i_n \leq D} \left[ \prod_{d=1}^n k_{i_d}(x_{i_d}, x'_{i_d}) \right] \quad (6.6)$$

$$k_{add_D}(\mathbf{x}, \mathbf{x}') = \sigma_D^2 \sum_{1 \leq i_1 < i_2 < \dots < i_D \leq D} \left[ \prod_{d=1}^D k_{i_d}(x_{i_d}, x'_{i_d}) \right] = \sigma_D^2 \prod_{d=1}^D k_d(x_d, x'_d) \quad (6.7)$$

where  $D$  is the dimension of our input space, and  $\sigma_n^2$  is the variance assigned to all  $n$ th order interactions. The  $n$ th covariance function is a sum of  $\binom{D}{n}$  terms. In particular, the  $D$ th order additive covariance function has  $\binom{D}{D} = 1$  term, a product of each dimension's covariance function. In the case where each base kernel is a one-dimensional squared-exponential kernel, the  $D$ th-order term corresponds to the multivariate squared-exponential kernel, also known as SE-ARD:

$$k_{add_D}(\mathbf{x}, \mathbf{x}') = \sigma_D^2 \prod_{d=1}^D k_d(x_d, x'_d) = \sigma_D^2 \prod_{d=1}^D \exp \left( -\frac{(x_d - x'_d)^2}{2l_d^2} \right) = \sigma_D^2 \exp \left( -\sum_{d=1}^D \frac{(x_d - x'_d)^2}{2l_d^2} \right) \quad (6.8)$$

also commonly known as the Gaussian kernel.

The full additive kernel is a sum of the additive kernels of all orders.

The only design choice necessary to specify an additive kernel is the selection of a one-dimensional base kernel for each input dimension. Parameters of the base kernels (such as length-scales) can be learned as usual by maximizing the marginal likelihood of the training data.

### 6.2.1 Weighting different orders of interaction

In addition to the parameters of each dimension's kernel, additive kernels are equipped with a set of  $D$  parameters  $\sigma_1^2 \dots \sigma_D^2$ . These “order variance” parameters have a useful

interpretation: the  $d$ th order variance hyperparameter controls how much of the target function's variance comes from interactions of the  $d$ th order. Table 6.1 shows examples of the variance contributed by different orders of interaction, learned on real datasets.

Table 6.1 Percentage of variance contributed by each order of the additive model, on different datasets. The maximum order of interaction is set to the input dimension or 10, whichever is smaller.

Dataset	Order of interaction									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
pima	0.1	0.1	0.1	0.3	1.5	<b>96.4</b>	1.4	0.0		
liver	0.0	0.2	<b>99.7</b>	0.1	0.0	0.0				
heart	<b>77.6</b>	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	22.0
concrete	<b>70.6</b>	13.3	13.8	2.3	0.0	0.0	0.0	0.0		
pumadyn-8nh	0.0	0.1	0.1	0.1	0.1	0.1	0.1	<b>99.5</b>		
servo	<b>58.7</b>	27.4	0.0	13.9						
housing	0.1	0.6	<b>80.6</b>	1.4	1.8	0.8	0.7	0.8	0.6	12.7

On different datasets, the dominant order of interaction estimated by the additive model varies widely. An additive GP with all of its variance coming from the 1st order is equivalent to a sum of one-dimensional functions. An additive GP with all its variance coming from the  $D$ th order is equivalent to a GP with an SE-ARD kernel.

Because the variance parameters can specify which degrees of interaction are important, the additive GP can capture many different types of structure. The marginal likelihood will usually favor using lower orders if possible, since the  $|K|^{-\frac{1}{2}}$  term in the GP marginal likelihood (equation (1.3)) will usually be larger for less flexible model classes. Low-order structure allows long-range extrapolation, as shown in section 2.4.2. If low-dimensional additive structure is not present, the kernel parameters can specify a suitably flexible model, with interactions between as many variables as necessary.

### 6.2.2 Efficiently evaluating additive kernels

An additive kernel over  $D$  inputs with interactions up to order  $n$  has  $O(2^n)$  terms. Naïvely summing over these terms quickly becomes intractable. Perhaps surprisingly, one can evaluate the sum over all terms in  $O(D^2)$ , while also weighting each order of interaction separately.

To efficiently compute the additive kernel, we exploit the fact that the  $n$ th order

additive kernel corresponds to the  $n$ th *elementary symmetric polynomial* (Macdonald, 1998) of the base kernels, which we denote  $e_n$ . For example: if  $\mathbf{x}$  has 4 input dimensions ( $D = 4$ ), and if we use the shorthand notation  $k_d = k_d(x_d, x'_d)$ , then

$$k_{\text{add}_0}(\mathbf{x}, \mathbf{x}') = e_0(k_1, k_2, k_3, k_4) = 1 \quad (6.9)$$

$$k_{\text{add}_1}(\mathbf{x}, \mathbf{x}') = e_1(k_1, k_2, k_3, k_4) = k_1 + k_2 + k_3 + k_4 \quad (6.10)$$

$$k_{\text{add}_2}(\mathbf{x}, \mathbf{x}') = e_2(k_1, k_2, k_3, k_4) = k_1 k_2 + k_1 k_3 + k_1 k_4 + k_2 k_3 + k_2 k_4 + k_3 k_4 \quad (6.11)$$

$$k_{\text{add}_3}(\mathbf{x}, \mathbf{x}') = e_3(k_1, k_2, k_3, k_4) = k_1 k_2 k_3 + k_1 k_2 k_4 + k_1 k_3 k_4 + k_2 k_3 k_4 \quad (6.12)$$

$$k_{\text{add}_4}(\mathbf{x}, \mathbf{x}') = e_4(k_1, k_2, k_3, k_4) = k_1 k_2 k_3 k_4 \quad (6.13)$$

The Newton-Girard formulae give an efficient recursive form for computing these polynomials:

$$k_{\text{add}_n}(\mathbf{x}, \mathbf{x}') = e_n(k_1, \dots, k_D) = \frac{1}{n} \sum_{a=1}^n (-1)^{(a-1)} e_{n-a}(k_1, \dots, k_D) \sum_{i=1}^D k_i^a \quad (6.14)$$

The Newton-Girard formulae have time complexity  $\mathcal{O}(D^2)$ , while computing a sum over an exponential number of terms.

### Evaluation of derivatives

Conveniently, we can use the same trick to efficiently compute all of the necessary derivatives of the additive kernel with respect to the base kernels. We merely need to remove the kernel of interest from each term of the polynomials:

$$\frac{\partial k_{\text{add}_n}}{\partial k_j} = e_{n-1}(k_1, \dots, k_{j-1}, k_{j+1}, \dots, k_D) \quad (6.15)$$

This allows us to optimize the base kernel parameters with respect to the marginal likelihood using gradient-based methods.

### Computational cost

The computational cost of evaluating the Gram matrix  $k(\mathbf{X}, \mathbf{X})$  of a product kernel such as the SE-ARD scales as  $\mathcal{O}(N^2 D)$ , while the cost of evaluating the Gram matrix of the additive kernel scales as  $\mathcal{O}(N^2 DR)$ , where  $R$  is the maximum degree of interaction allowed (up to  $D$ ). In higher dimensions, this can be a significant cost, even relative to the fixed  $\mathcal{O}(N^3)$  cost of inverting the Gram matrix. However, table 6.1 shows that

sometimes only the first few orders of interaction are important. Hence if one is computationally limited, one may be able to limit the maximum degree of interaction without losing much accuracy.

### 6.3 Additive models allow non-local interactions

Popular kernels such as the SE, RQ or Matérn kernels are *local* kernels, depending only on the scaled Euclidean distance between two points, having the form:

$$k(\mathbf{x}, \mathbf{x}') = g\left(\sum_{d=1}^D \left(\frac{x_d - x'_d}{l_d}\right)^2\right) \quad (6.16)$$

For some function  $g(\cdot)$ . Bengio et al. (2006) argue that models based on local kernels are particularly susceptible to the curse of dimensionality, and are generally unable to extrapolate away from the training data. Thus, methods based solely on local kernels will sometimes require training examples at exponentially-many combinations of inputs.

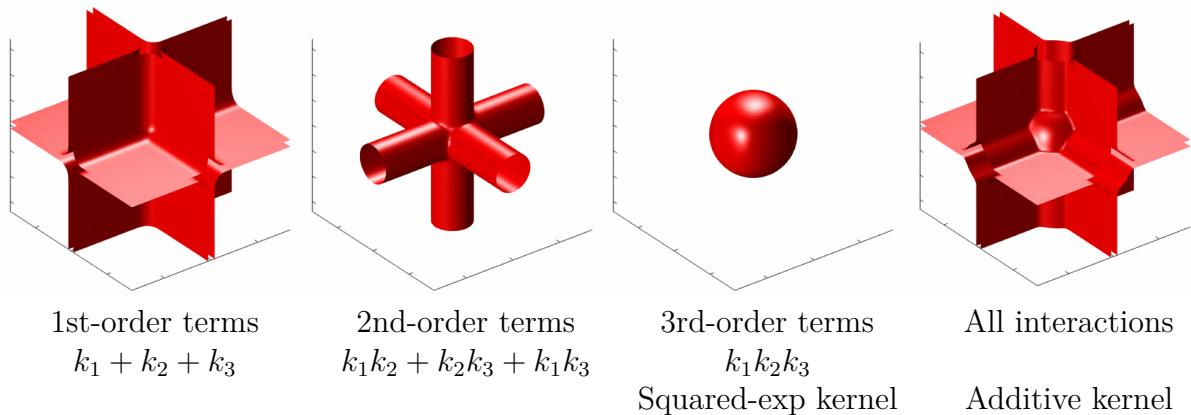


Figure 6.1 Isocontours of additive kernels in 3 dimensions. The third-order kernel only considers nearby points relevant, while the lower-order kernels allow the output to depend on distant points, as long as they share one or more input value.

Additive kernels have a more complex structure, and allow extrapolation far from the training data. For example, additive kernels of the second order give high covariance between function values at points which are similar in any two input dimensions. Figure 6.1 provides a geometric comparison between squared-exponential kernels and additive kernels in 3 dimensions. Section 2.4.1 contains an example of how additive kernels extrapolate differently than local kernels.

## 6.4 Dropout in Gaussian processes

*Dropout* is a method for regularizing neural networks (Hinton et al., 2012; Srivastava, 2013). Training with dropout entails randomly and independently setting to zero (“dropping”) some proportion  $p$  of features or inputs, in order to improve the robustness of the resulting network by reducing co-dependence between neurons. To maintain similar overall activation levels, weights are multiplied by  $1/p$  at test time. Alternatively, feature activations are multiplied by  $1/p$  during training. Test-time predictions are made by approximately averaging over all possible ways of dropping out neurons.

Baldi and Sadowski (2013) and Wang and Manning (2013) analyzed dropout in terms of the effective prior induced by this procedure in several models, such as linear and logistic regression. In this section, we perform a similar analysis for GPs, examining the priors on functions that result from performing dropout in the one-hidden-layer neural network implicitly defined by a GP.

Recall from section 5.1 that GPs can be derived as an infinitely-wide one-layer neural network, with fixed activation functions  $\mathbf{h}(\mathbf{x})$  (where  $k(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}')$ ), and independent random weights  $\boldsymbol{\alpha}$  with finite variance  $\sigma_\alpha^2$ :

$$f(\mathbf{x}) = \frac{1}{K} \boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^K \alpha_i h_i(\mathbf{x}) \quad (6.17)$$

$$\implies f \xrightarrow{K \rightarrow \infty} \mathcal{GP}\left(\mathbb{E}[\boldsymbol{\alpha}]^\top \mathbf{h}(\mathbf{x}), \sigma_\alpha^2 \mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}')\right) \quad (6.18)$$

Mercer’s theorem implies that we can write any GP prior equivalently in this way. Having expressed a GP as a neural network, we can examine the prior we get from performing dropout in this network.

This result does not hold in neural networks having a finite number of hidden features with Gaussian-distributed weights, another model class that gives rise to GPs.

### 6.4.1 Dropout on hidden layers

First, we will examine the prior we get from independently dropping features from  $\mathbf{h}(\mathbf{x})$  by setting some of the weights  $\boldsymbol{\alpha}$  to zero with probability  $p$ . For simplicity, we assume that  $\mathbb{E}[\boldsymbol{\alpha}] = \mathbf{0}$ . If the weights initially have finite variance  $\sigma_\alpha^2$  before dropout, then after dropout they’ll have variance

$$r_i \stackrel{\text{iid}}{\sim} \text{Ber}(p) \quad \mathbb{V}[r_i \alpha_i] = p \sigma_\alpha^2. \quad (6.19)$$

Because equation (6.18) is a result of the central limit theorem, it does not depend on the form of the distribution on  $\alpha$ , only its mean and variance. Thus, dropping out features of an infinitely-wide MLP does not change the model at all, except to rescale the output variance. Indeed, multiplying all weights by  $p^{-1/2}$  restores the initial variance:

$$\mathbb{V} \left[ \frac{1}{p^{\frac{1}{2}}} r_i \alpha_i \right] = \frac{p}{p} \sigma_{\alpha}^2 = \sigma_{\alpha}^2. \quad (6.20)$$

In which case dropout on the hidden units has no effect at all. Intuitively, this is because no individual feature can have more than an infinitesimal contribution to the network output.

### 6.4.2 Dropout on inputs gives additive covariance

We can also perform dropout on the  $D$  inputs of the GP. For simplicity, we'll consider a stationary product kernel  $k(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^D k_d(x_d, x'_d)$  which has been normalized such that  $k(\mathbf{x}, \mathbf{x}') = 1$ , and dropout probability of  $1/2$ . In this case, the generative model can be written as:

$$r_i \stackrel{\text{iid}}{\sim} \text{Ber} \left( \frac{1}{2} \right), \quad f(\mathbf{x}) | \mathbf{r} \sim \mathcal{GP} \left( 0, \prod_{d=1}^D k_d(x_d, x'_d)^{r_d} \right) \quad (6.21)$$

This is a mixture of GPs, each depending on a different subset of the inputs:

$$p(f(\mathbf{x})) = \sum_{\mathbf{r}} p(f(\mathbf{x}) | \mathbf{r}) p(\mathbf{r}) = \frac{1}{2^D} \sum_{\mathbf{r} \in \{0,1\}^D} \mathcal{GP} \left( f(\mathbf{x}) \mid 0, \prod_{d=1}^D k_d(x_d, x'_d)^{r_d} \right) \quad (6.22)$$

We present two results ways to gain intuition about this model.

First, if the kernel on each dimension has the form  $k_d(x_d, x'_d) = g \left( \frac{x_d - x'_d}{w_d} \right)$ , as does the SE kernel, then any input dimension can be dropped out by setting its lengthscale  $w_d$  to  $\infty$ . Thus, performing dropout on the inputs of a GP corresponds to putting independent spike-and-slab priors on the lengthscales, with each dimension's distribution independently having “spikes” at  $w_d = \infty$  with probability mass of  $1/2$ .

Another way to understand the resulting prior is to note that the dropout mixture (equation (6.22)) has the same covariance as an additive GP, scaled by a factor of  $2^{-D}$ :

$$\text{cov} \left( \begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}') \end{bmatrix} \right) = \frac{1}{2^D} \sum_{\mathbf{r} \in \{0,1\}^D} \prod_{d=1}^D k_d(x_d, x'_d)^{r_d} \quad (6.23)$$

Therefore, ignoring higher moments, dropout on the inputs of a GP can be approximated by an additive GP with all orders equally weighted. For dropout rates  $p$  other than  $1/2$ , the  $d$  order terms will be weighted by the corresponding binomial coefficient  $\binom{D}{d}$ . This suggests an interpretation of additive GPs as an approximation to a mixture of models where each model only depends on a subset of the input variables.

## 6.5 Related work

Since additive models are a relatively natural and easy-to-analyze model class, the literature on similar model classes is extensive. This section attempts to provide a broad overview.

### Previous examples of additive GPs

Since the non-local structure capturable by additive kernels is necessarily axis-aligned, we can naturally consider that an initial transformation of the input space might allow us to recover non-axis aligned additivity in functions. This avenue was explored by [Gilboa et al. \(2013\)](#), who developed a linearly-transformed first-order additive GP model, called projection-pursuit GP regression. They further showed that inference in this model was possible in  $\mathcal{O}(N)$  time.

[Durrande et al. \(2011\)](#) also examined the properties of additive GPs, and proposed a layer-wise optimization strategy for kernel hyperparameters in these models.

[Plate \(1999\)](#) constructed an additive GP having only first-order and  $D$ th-order terms. This model is motivated by the desire to trade off the interpretability of first-order models with the flexibility of full-order models. Our experiments show that sometimes, the intermediate degrees of interaction contribute most of the variance.

[Kaufman and Cain \(2010\)](#) used a closely related procedure called Gaussian process ANOVA to perform a Bayesian analysis of meteorological data using 2nd and 3rd-order interactions. They also introduce a weighting scheme to ensure that each order's total contribution sums to zero. It is not clear if this weighting scheme permits the use of the Newton-Girard formula to speed computation of the Gram matrix.

### Hierarchical kernel learning

A similar model class was recently explored by [Bach \(2009b\)](#) called hierarchical kernel learning (HKL). HKL uses a regularized optimization framework to learn a weighted

sum over an exponential number of kernels which can be computed in polynomial time. This method chooses among a *hull* of kernels, defined as a set of terms such that if  $\prod_{j \in J} k_j(\mathbf{x}, \mathbf{x}')$  is included in the set, then so are all lower-order terms containing the same elements:  $\prod_{j \in J/i} k_j(\mathbf{x}, \mathbf{x}')$ , for all  $i \in J$ . HKL computes the sum over all orders in  $\mathcal{O}(D)$  time by the formula:

$$k_a(\mathbf{x}, \mathbf{x}') = v^2 \prod_{d=1}^D (1 + \alpha k_d(x_d, x'_d)) \quad (6.24)$$

which forces the weight of all  $n$ th order terms to be weighted by  $\alpha^n$ .

Figure 6.2 contrasts the HKL model class with the additive GP model. Neither method is strictly more flexible than the other. The main difficulty with this approach is that the kernel parameters are hard to set, other than by cross-validation.

## Support vector machines

[Vapnik \(1998\)](#) introduced the support vector ANOVA decomposition, which has the same form as our additive kernel. They recommend approximating the sum over all interactions with only one of the  $D$  sets of interactions “of appropriate order”, presumably because of the difficulty of setting the parameters of an SVM. This is an example of a model choice which can be automated in the GP framework.

[Stitson et al. \(1999\)](#) performed experiments which favourably compared the predictive accuracy of the support vector ANOVA decomposition against polynomial and spline kernels. They too allowed only one order to be active, and set parameters by cross-validation.

## Other related models

A closely related procedure from [Wahba \(1990\)](#) is smoothing-splines ANOVA (SS-ANOVA). An SS-ANOVA model is a weighted sum of splines along each dimension, plus a sum of splines over all pairs of dimensions, all triplets, etc, with each individual interaction term having a separate weighting parameter. Because the number of terms to consider grows exponentially in the order, in practice, only terms of first and second order are usually considered.

This more general model class, in which each interaction term is estimated separately, is known in the physical sciences as High Dimensional Model Representation (HDMR). [Rabitz and Aliş \(1999\)](#) review some properties and applications of this model class.

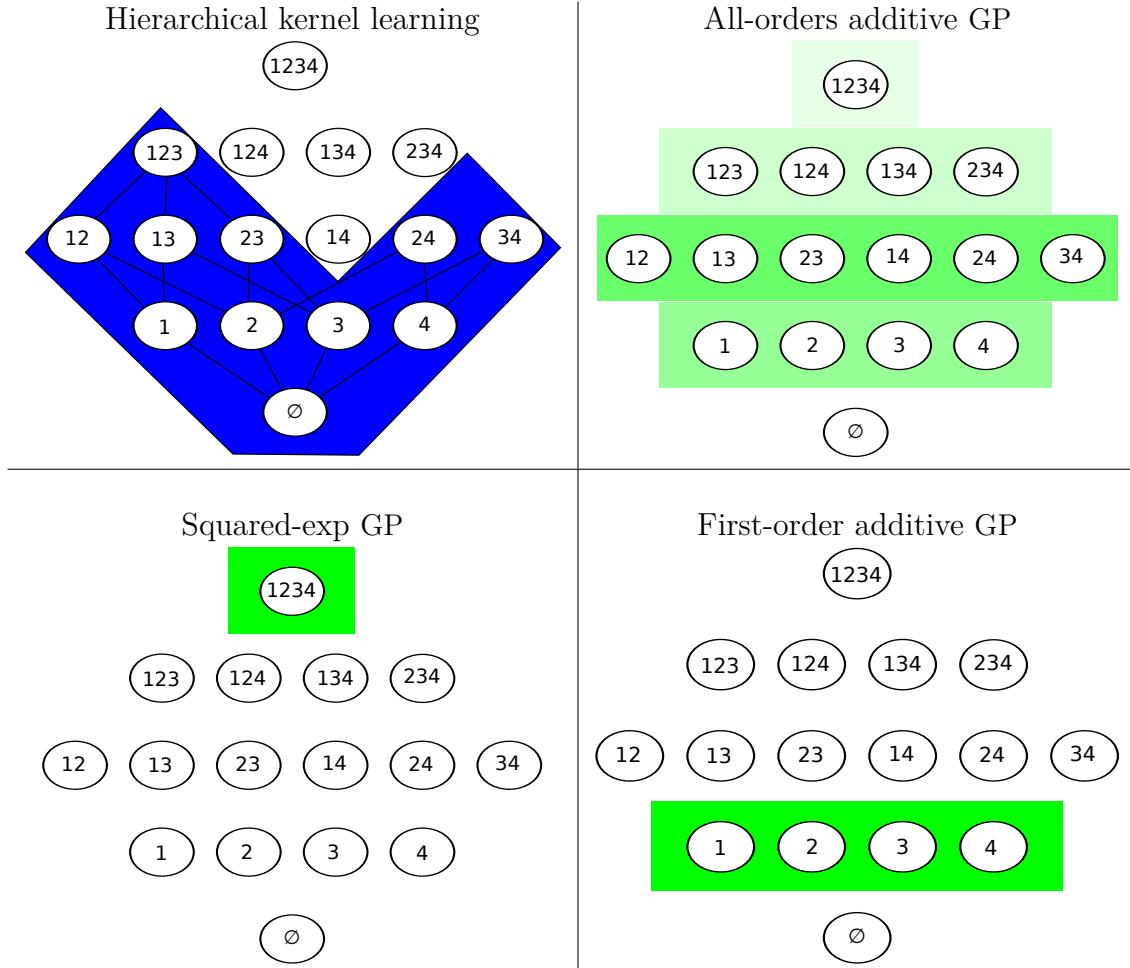


Figure 6.2 A comparison of different additive model classes. Nodes represent different interaction terms, ranging from first-order to fourth-order interactions. Coloured boxes represent the weightings of different terms. *Top left:* HKL can select a hull of interaction terms, but must use a pre-determined weighting over those terms. *Top right:* the additive GP model can weight each order of interaction separately, but weights all terms equally within each order. *Bottom row:* The SE-GP and first-order additive GP models are special cases of the all-orders additive GP.

The main benefits of the model setup and parameterization proposed in this chapter are the ability to include all  $D$  orders of interaction with differing weights, and the ability to learn kernel parameters individually per input dimension, allowing automatic relevance determination to operate.

## 6.6 Regression and classification experiments

### Choosing the Base Kernel

A  $D$ -dimensional SE-ARD kernel has  $D$  lengthscale parameters and one output variance parameter. An first-order additive SE model has  $D$  lengthscale parameters and one  $D$  output variance parameters. A fully-parametrized model including all orders of interaction with a separate output variance for each scale will have  $3 \times D - 1$  effective parameters. Because each additional parameter increases the tendency to overfit, in our experiments we fixed each one-dimensional kernel's output variance to be 1, and only learned the length-scale of each kernel.

### Methods

We compare six different methods. In the results tables below, GP Additive refers to a GP using the additive kernel with squared-exp base kernels. For speed, we limited the maximum order of interaction to 10. GP-1st denotes an additive GP model with only first-order interactions - a sum of one-dimensional kernels. GP Squared-exp is a GP model with a SE-ARD kernel. HKL was run using the all-subsets kernel, which corresponds to the same set of interaction terms as considered by the additive GP with a squared-exp base kernel.

For all GP models, we fit kernel parameters by the standard method of maximizing training-set marginal likelihood, using L-BFGS ([Nocedal, 1980](#)) for 500 iterations, allowing five random restarts. In addition to learning kernel parameters, we fit a constant mean function to the data. In the classification experiments, approximate GP inference was done using expectation propagation ([Minka, 2001](#)).

For the regression experiments, we also compared against the structure search method from chapter 3, run up to depth 10, using the SE and RQ base kernel families.

### 6.6.1 Datasets

We compared these methods on a diverse set of regression and classification datasets from the UCI repository (Bache and Lichman, 2013). Their size and dimension are given in tables 6.2 and 6.3:

Table 6.2 Regression Dataset Statistics

Method	bach	concrete	pumadyn	servo	housing
Dimension	8	8	8	4	13
Number of datapoints	200	500	512	167	506

Table 6.3 Classification Dataset Statistics

Method	breast	pima	sonar	ionosphere	liver	heart
Dimension	9	8	60	32	6	13
Number of datapoints	449	768	208	351	345	297

#### Bach synthetic dataset

In addition to standard UCI repository datasets, we generated a synthetic dataset following the same recipe as Bach (2009b). This dataset was designed to demonstrate the advantages of HKL over GP-SE. It is generated by passing correlated Gaussian-distributed inputs  $x_1, x_2, \dots, x_8$  through the quadratic function

$$f(\mathbf{x}) = \sum_{i=1}^4 \sum_{j=i+1}^4 x_i x_j + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon) \quad (6.25)$$

This dataset will presumably be well-modeled by an additive kernel which includes all two-way interactions over the first 4 variables, but does not depend on the extra 4 correlated nuisance inputs or the higher-order interactions.

### 6.6.2 Results

Tables 6.4 to 6.7 show mean performance across 10 train-test splits. Because HKL does not specify a noise model, it could not be included in the likelihood comparisons.

Table 6.4 Regression mean squared error

Method	bach	concrete	pumadyn-8nh	servo	housing
Linear Regression	1.031	0.404	0.641	0.523	0.289
GP-1st	1.259	0.149	0.598	0.281	0.161
HKL	<b>0.199</b>	0.147	0.346	0.199	0.151
GP Squared-exp	<b>0.045</b>	0.157	0.317	0.126	<b>0.092</b>
GP Additive	<b>0.045</b>	<b>0.089</b>	<b>0.316</b>	<b>0.110</b>	0.102
Structure Search	<b>0.044</b>	<b>0.087</b>	<b>0.315</b>	<b>0.102</b>	<b>0.082</b>

Table 6.5 Regression negative log-likelihood

Method	bach	concrete	pumadyn-8nh	servo	housing
Linear Regression	2.430	1.403	1.881	1.678	1.052
GP-1st	1.708	0.467	1.195	0.800	0.457
GP Squared-exp	<b>-0.131</b>	0.398	0.843	0.429	0.207
GP Additive	<b>-0.131</b>	<b>0.114</b>	<b>0.841</b>	<b>0.309</b>	0.194
Structure Search	<b>-0.141</b>	<b>0.065</b>	<b>0.840</b>	<b>0.265</b>	<b>0.059</b>

Table 6.6 Classification percent error

Method	breast	pima	sonar	ionosphere	liver	heart
Logistic Regression	7.611	24.392	26.786	16.810	45.060	<b>16.082</b>
GP-1st	<b>5.189</b>	<b>22.419</b>	<b>15.786</b>	<b>8.524</b>	<b>29.842</b>	<b>16.839</b>
HKL	<b>5.377</b>	24.261	<b>21.000</b>	9.119	<b>27.270</b>	<b>18.975</b>
GP Squared-exp	<b>4.734</b>	<b>23.722</b>	<b>16.357</b>	<b>6.833</b>	<b>31.237</b>	<b>20.642</b>
GP Additive	<b>5.566</b>	<b>23.076</b>	<b>15.714</b>	<b>7.976</b>	<b>30.060</b>	<b>18.496</b>

Table 6.7 Classification negative log-likelihood

Method	breast	pima	sonar	ionosphere	liver	heart
Logistic Regression	0.247	0.560	4.609	0.878	0.864	0.575
GP-1st	<b>0.163</b>	<b>0.461</b>	<b>0.377</b>	<b>0.312</b>	<b>0.569</b>	<b>0.393</b>
GP Squared-exp	<b>0.146</b>	0.478	<b>0.425</b>	<b>0.236</b>	<b>0.601</b>	0.480
GP Additive	<b>0.150</b>	<b>0.466</b>	<b>0.409</b>	<b>0.295</b>	<b>0.588</b>	<b>0.415</b>

The model with best performance on each dataset is in bold, along with all other models that were not significantly different under a paired  $t$ -test. The additive and structure search methods usually outperformed the other methods, especially on regression problems.

The structure search outperforms the additive GP, but at the cost of a slow search over kernels. The additive GP performed best on datasets well-explained by low orders of interaction, and approximately as well as the SE-GP model on datasets which were well explained by high orders of interaction (see table 6.1). Because the additive GP is a superset of both the GP-1st model and the SE-GP model, instances where the additive GP performs slightly worse are presumably due to over-fitting, or due to the hyperparameter optimization becoming stuck in a local maximum. Performance could be expected to benefit from approximately integrating over the kernel parameters.

The performance of HKL is consistent with the results in Bach (2009b), performing competitively but slightly worse than SE-GP.

### Source code

Additive Gaussian processes are particularly appealing in practice because their use requires only the specification of the base kernel; all other aspects of GP inference remain the same. Note that we are also free to choose a different covariance function along each dimension.

All of the experiments in this chapter were performed using the standard GPML toolbox, available at [gaussianprocess.org/gpml/code](http://gaussianprocess.org/gpml/code). The additive kernel described in this chapter is included in the latest release. Code to perform all experiments in this chapter is available at [github.com/duvenaud/additive-gps](https://github.com/duvenaud/additive-gps)

## 6.7 Conclusions

In this chapter, we presented a tractable GP model consisting of a sum of exponentially-many functions, each depending on a different subset of the inputs. Our experiments indicate that, to varying degrees, such additive structure is useful for modeling real datasets. When it is present, modeling this structure allows our model to perform better than standard GP models. In the case where no such structure exists, the higher-order interaction terms present in the kernel can recover arbitrarily flexible models, as well. The additive GP also affords some degree of interpretability: the variance parameters on each order of interaction indicate which sorts of structure are present in the data.

The model class considered in this chapter is a subset of that explored by the structure search presented in chapter 3. Thus additive GPs can be considered a quick-and-dirty structure search, being strictly more limited in the types of structure that it can discover, but much faster and simpler to implement.

Related model classes have been previously explored, most notably smoothing-splines ANOVA, and the support vector ANOVA decomposition. However, these models are difficult to apply in practice because kernel parameters, regularization penalties, and the relevant orders of interaction must all be set by hand or by cross-validation. This chapter illustrates that the GP framework allows these model choices to be performed automatically.

# Chapter 7

## Warped Mixture Models

“What, exactly, is a cluster?”

- Bernhard Schölkopf, personal communication

Previous chapters showed how the probabilistic nature of GPs lets us automatically include an appropriate amount of structure, and the correct kind, when building models of functions. In this chapter, we show how to take advantage of this property when composing GPs with other models, automatically balancing between the complexity of the GP and the other parts of the model.

This chapter considers a simple example: a Gaussian mixture model, warped by a draw from a GP. This novel model produces clusters with arbitrary shapes, depending on the warping. We call the proposed model the *infinite warped mixture model* (iWMM). Figure 7.3 shows a set of manifolds and datapoints sampled from the prior defined by this model. The probabilistic nature of the iWMM lets us automatically infer the number, dimension, and shape of a set of nonlinear manifolds, and summarize those manifolds in a low-dimensional latent space.

The work comprising the bulk of this chapter was done in collaboration with Tomoharu Iwata and Zoubin Ghahramani, and appeared in [Iwata et al. \(2013\)](#). The main idea was born out of a conversation between Tomo and myself, and together we wrote almost all of the code as well as the paper. Tomoharu ran most of the experiments. Zoubin Ghahramani provided guidance and many helpful suggestions throughout the project.

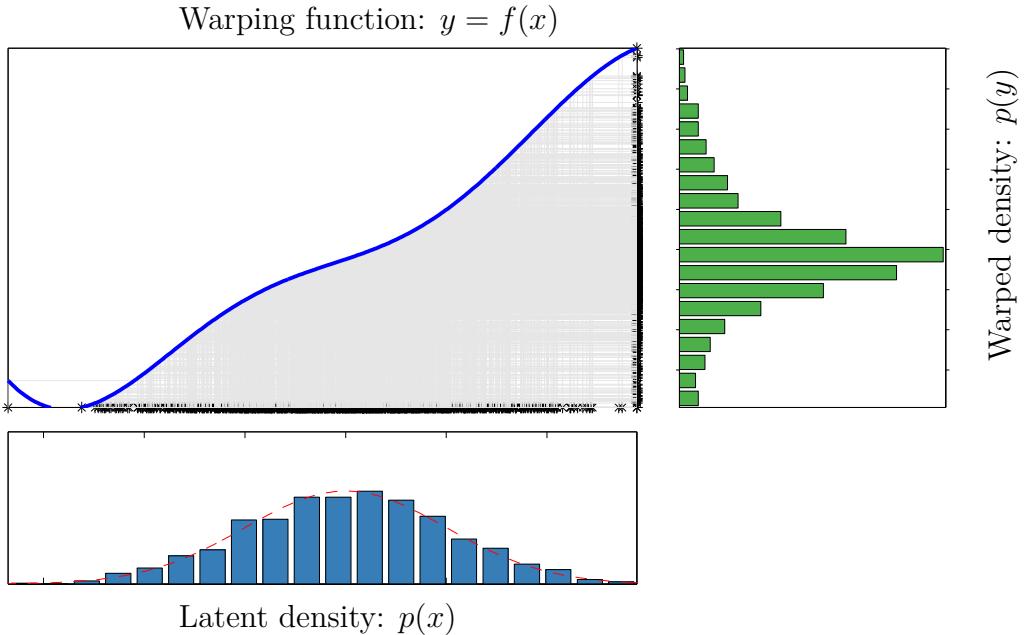


Figure 7.1 A draw from a one-dimensional Gaussian process latent variable model. *Bottom:* the density of a set of samples from a 1D Gaussian, specifying the distribution  $p(x)$  in the latent space. *Top Right:* A function  $y = f(x)$  drawn from a GP prior. Grey lines show points being mapped through  $f$ . *Left:* A nonparametric density  $p(y)$  defined by warping the latent density through the sampled function.

## 7.1 The Gaussian process latent variable model

The iWMM can be viewed as an extension of the Gaussian process latent variable model (GP-LVM) (Lawrence, 2004), a probabilistic model of nonlinear manifolds. The GP-LVM smoothly warps a Gaussian density into a more complicated distribution, using a draw from a GP. Usually, we think of the Gaussian density as living in a “latent space” having  $Q$  dimensions, and the warped density living in the observed space having  $D$  dimensions.

A standard definition of the GP-LVM is as follows:

$$\text{latent coordinates } \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)^T \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{x}|0, \mathbf{I}_Q) \quad (7.1)$$

$$\text{warping functions } \mathbf{f} = (f_1, f_2, \dots, f_D)^T \stackrel{\text{iid}}{\sim} \mathcal{GP}(0, \text{SE} + \text{WN}) \quad (7.2)$$

$$\text{observed datapoints } \mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)^T = \mathbf{f}(\mathbf{X}) \quad (7.3)$$

Under the GP-LVM, the probability of observations given the latent coordinates

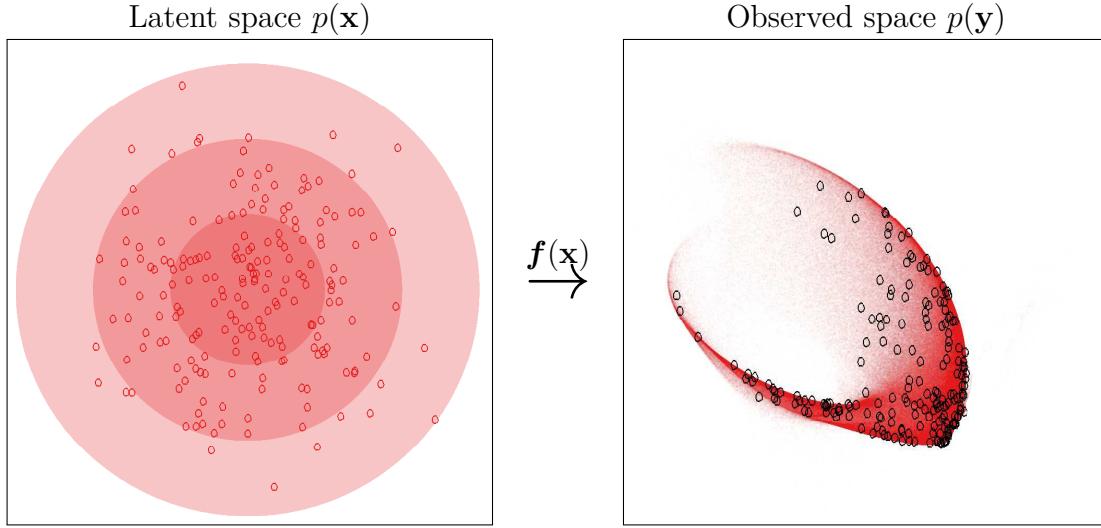


Figure 7.2 A draw from a multi-dimensional Gaussian process latent variable model. *Left:* Isocontours and samples from a 2D Gaussian, specifying the distribution  $p(\mathbf{x})$  in the latent space. *Right:* Observed density  $p(\mathbf{y})$  has a nonparametric shape, defined by warping the latent density through a function drawn from a GP prior.

integrating out the mapping functions, is simply a product of multivariate normals:

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{d=1}^D p(\mathbf{Y}_{:,d}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{d=1}^D \mathcal{N}(\mathbf{Y}_{:,d}|0, \mathbf{K}_{\boldsymbol{\theta}}) \quad (7.4)$$

$$= (2\pi)^{-\frac{DN}{2}} |\mathbf{K}|^{-\frac{D}{2}} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{Y}^T \mathbf{K}^{-1} \mathbf{Y})\right), \quad (7.5)$$

where  $\boldsymbol{\theta}$  are the kernel parameters and  $\mathbf{K}$  is the Gram matrix  $k(\mathbf{X}, \mathbf{X})$ .

Typically, the GP-LVM is used for dimensionality reduction or visualization, and the latent coordinates are set by maximizing (7.5). In that setting, the Gaussian prior density on  $\mathbf{x}$  is essentially a regularizer which keeps the latent coordinates from spreading arbitrarily far apart. In contrast, we integrate out the latent coordinates, and the iWMM places a more flexible parameterization on  $p(\mathbf{x})$  than a single isotropic Gaussian. Just as the GP-LVM can be viewed as a manifold learning algorithm, the iWMM can be viewed as learning a set of manifolds, one for each cluster.

## 7.2 The infinite warped mixture model

This section defines in detail the infinite warped mixture model (iWMM). Like the GP-LVM, the iWMM assumes a smooth nonlinear mapping from a latent density to an

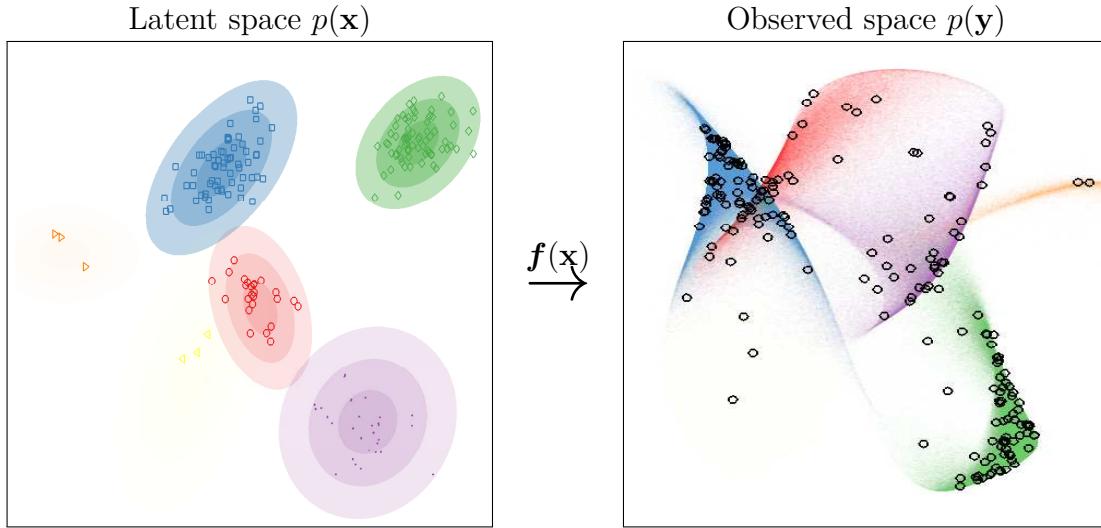


Figure 7.3 A sample from the iWMM prior. *Left:* In the latent space, a mixture distribution is sampled from a Dirichlet process mixture of Gaussians. *Right:* The latent mixture is smoothly warped to produce a set of non-Gaussian manifolds in the observed space.

observed density. The difference is that the iWMM assumes that the latent density is an infinite Gaussian mixture model (iGMM) (Rasmussen, 2000):

$$p(\mathbf{x}) = \sum_{c=1}^{\infty} \lambda_c \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_c, \mathbf{R}_c^{-1}) \quad (7.6)$$

where  $\lambda_c$ ,  $\boldsymbol{\mu}_c$  and  $\mathbf{R}_c$  is the mixture weight, mean, and precision matrix of the  $c^{\text{th}}$  mixture component.

The iWMM can be seen as a generalization of either the GP-LVM or the iGMM. To be precise, the iWMM with a single fixed spherical Gaussian density on the latent coordinates corresponds to the GP-LVM, while the iWMM with fixed direct mapping function  $f_d(\mathbf{x}) = x_d$  and  $Q = D$  corresponds to the iGMM.

A flexible model of cluster shapes is required to correctly estimate the number of clusters, if those clusters do not happen to be Gaussian. For example, a mixture of Gaussians fit to a single non-Gaussian cluster (such as one that is curved or heavy-tailed) will report that the data contains many Gaussian clusters.

## 7.3 Inference

As discussed in section 1.1.3, one of the main advantages of GP priors is that, given inputs  $\mathbf{X}$ , outputs  $\mathbf{Y}$  and kernel parameters  $\boldsymbol{\theta}$ , we can analytically integrate over functions mapping  $\mathbf{X}$  to  $\mathbf{Y}$ . However, inference becomes more difficult if we introduce uncertainty about the kernel parameters, or the input locations  $\mathbf{X}$ . This section outlines how to infer all parameters in the iWMM given only a set of observations  $\mathbf{Y}$ . Details can be found in chapter E.

By placing conjugate priors on the parameters of the Gaussian mixture components, we can analytically integrate out the cluster shapes, given the assignments of points to clusters. The only remaining variables to infer are the latent points  $\mathbf{X}$ , the cluster assignments  $\mathbf{Z}$ , and the kernel parameters  $\boldsymbol{\theta}$ . We can obtain samples from their posterior  $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta} | \mathbf{Y})$  by iterating two steps:

1. Given the latent points  $\mathbf{X}$ , we sample the discrete cluster memberships  $\mathbf{Z}$  using collapsed Gibbs sampling, integrating out the mixture parameters (E.9).
2. Given the cluster assignments  $\mathbf{Z}$ , we sample the continuous latent coordinates  $\mathbf{X}$  and kernel parameters  $\boldsymbol{\theta}$  using Hamiltonian Monte Carlo (HMC) (MacKay, 2003, chapter 30). The relevant equations are given by equations (E.11) to (E.14).

The complexity of each iteration of HMC is dominated by the  $\mathcal{O}(N^3)$  computation of  $\mathbf{K}^{-1}$ . This complexity could be improved by making use of an inducing-point approximation (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006).

### Posterior predictive density

One disadvantage of this model class is that its predictive density has no closed form. To approximate the predictive density, we sample latent points from the posterior on the latent density, and map them through warpings drawn from the corresponding posterior density. The Gaussian noise added to each observation means that each sample adds a Gaussian to the Monte Carlo estimate of the predictive density. Details can be found in chapter E. This procedure was used to generate the plots of posterior density in figures 7.3, 7.4 and 7.6.

## 7.4 Related work

The literature on manifold learning, clustering and dimensionality reduction is extensive. This section highlights some of the most relevant related work.

### Extensions of the GP-LVM

The GP-LVM has been used effectively in a wide variety of applications (Lawrence, 2004; Lawrence and Urtasun, 2009; Salzmann et al., 2008). The latent positions  $\mathbf{X}$  in the GP-LVM are typically obtained by maximum a posteriori estimation or variational Bayesian inference (Titsias and Lawrence, 2010), placing a single fixed spherical Gaussian prior on  $\mathbf{x}$ .

A regularized extension of the GP-LVM which allows estimation of the dimension of the latent space was introduced by Geiger et al. (2009), in which the latent variables and their intrinsic dimensionality are simultaneously optimized. The iWMM can also infer the intrinsic dimensionality of nonlinear manifolds: inferring the Gaussian covariance for each latent cluster allows the variance of irrelevant dimensions to become small. The marginal likelihood of the latent Gaussian mixture will favor using as few dimensions as possible to describe each cluster. In fact, because each latent cluster has a different set of parameters, each cluster can have a different effective dimension. This allows the iWMM to model manifolds of differing dimension in the observed space, as demonstrated in figure 7.4(b).

Nickisch and Rasmussen (2010) considered several modifications of the GP-LVM which model the latent density using a mixture of Gaussians centered around the latent points. They approximated the observed density  $p(\mathbf{Y})$  by another mixture of Gaussians, obtained by moment-matching the latent Gaussians after they had been warped into the observed space. Training was done by maximizing a leave-some-out predictive density. This method had poor predictive performance compared to simple baselines, and was not have a generative clustering model.

### Related linear models

The iWMM can also be viewed as a generalization of the mixture of probabilistic principle component analyzers (Tipping and Bishop, 1999), or the mixture of factor analyzers (Ghahramani and Beal, 2000), where the linear mapping of the mixtures is generalized to a nonlinear mapping by Gaussian processes, and the number of components is infinite.

### Non-probabilistic methods

There exist non-probabilistic clustering methods which can find clusters with complex shapes, such as spectral clustering (Ng et al., 2002) and nonlinear manifold clustering (Cao and Haralick, 2006; Elhamifar and Vidal, 2011). Spectral clustering finds clusters by first forming a similarity graph, then finding a low-dimensional latent representation using the graph, and finally, clustering the latent coordinates via k-means. The performance of spectral clustering depends on parameters which are usually set manually, such as the number of clusters, the number of neighbors, and the variance parameter used for constructing the similarity graph. The iWMM infers such parameters automatically, and has no need to construct a similarity graph.

The kernel Gaussian mixture model (Wang et al., 2003) can also find non-Gaussian shaped clusters. This model estimates a GMM in the implicit infinite-dimensional feature space defined by the kernel mapping of the observed space. However, the kernel parameters must be set by cross-validation. In contrast, the iWMM infers the mapping function such that the latent coordinates will be well-modeled by a mixture of Gaussians.

### Nonparametric cluster shapes

To the best of our knowledge, the only other Bayesian clustering method with nonparametric cluster shapes is that of Rodríguez and Walker (2012), who for one-dimensional data introduce a nonparametric model of *unimodal* clusters, where each cluster's density function strictly decreases away from its mode.

### Deep Gaussian processes

An elegant way to construct a GP-LVM with a more structured latent density  $p(\mathbf{x})$  is to use another GP-LVM to model the latent coordinates  $\mathbf{X}$ . This latent GP-LVM can have another GP-LVM modeling its latent space, etc. This is exactly the model class considered by Damianou and Lawrence (2013), who also test to what extent each layer's latent representation implicitly forms clusters. They found that when modeling MNIST hand-written digits, nearest-neighbour classification performed best in the 4th layer of a 5-layer deep nested GP-LVM, suggesting that the latent density might have been implicitly forming clusters at that level.

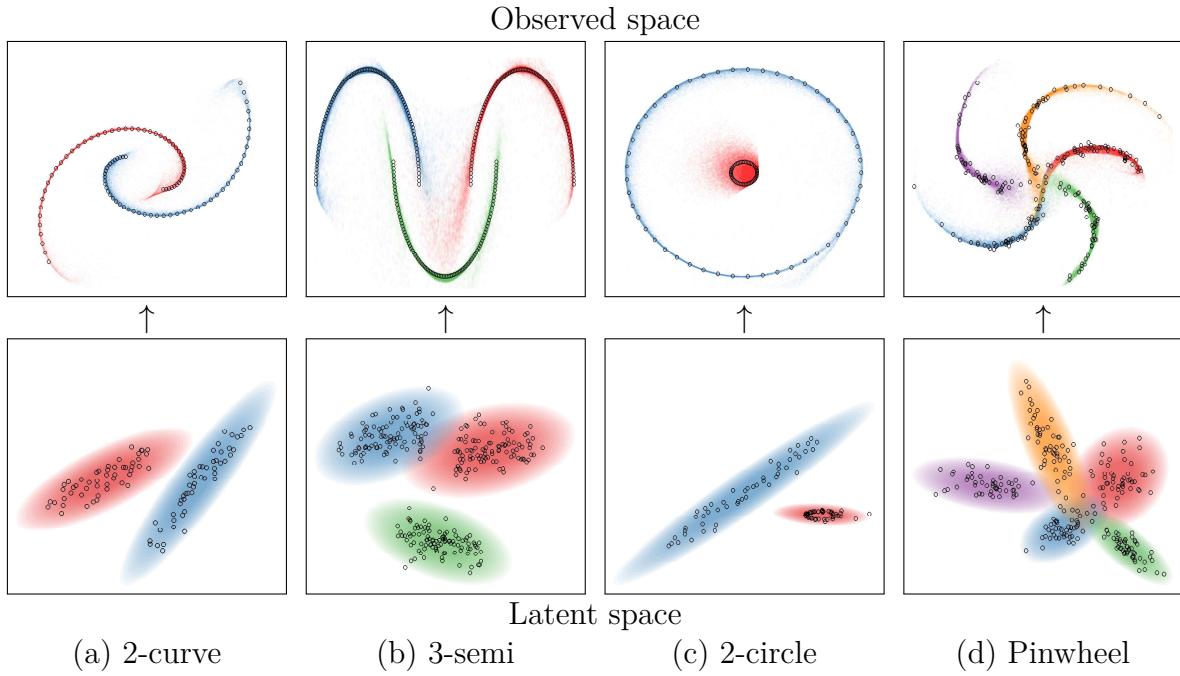


Figure 7.4 *Top row*: The observed unlabeled data points (black), and the cluster densities inferred by the iWMM (colors). *Bottom row*: Latent coordinates and Gaussian components from a single sample from the posterior. Each point in the latent space corresponds to a point in the observed space.

## 7.5 Experimental results

### 7.5.1 Synthetic datasets

Figure 7.4 demonstrates the proposed model on four synthetic datasets. None of these four datasets can be appropriately clustered by Gaussian mixture models (GMM). For example, consider the 2-curve data shown in Figure 7.4(a), where 100 data points lie in each of two curved lines in a two-dimensional observed space. A GMM with two components cannot separate the two curved lines, while a GMM with many components could separate the two lines only by breaking each line into many clusters. In contrast, the iWMM represents the two non-Gaussian clusters in the observed space by two Gaussian-shaped clusters in the latent space. Figure 7.4(b) shows a simple extension to three non-linearly-separable clusters.

Figure 7.4(c) shows an interesting manifold learning challenge: a dataset consisting of two concentric circles. The outer circle is modeled in the latent space by a Gaussian with one effective degree of freedom. This linear topology is fit to the outer circle in the

observed space by bending the two ends until they cross over. In contrast, the sampler fails to discover the 1D topology of the inner circle, modeling it with a 2D manifold instead. This example demonstrates that each cluster in the iWMM manifold can have a different effective dimension.

Figure 7.4(d) shows a five-armed variant of the pinwheel dataset of Adams and Ghahramani (2009), generated by warping a mixture of Gaussians into a spiral. This generative process closely matches the assumptions of the iWMM. Unsurprisingly, the iWMM is able to recover an analogous latent structure, and its predictive density follows the observed data manifolds.

### 7.5.2 Clustering face images

We also examined our model’s ability to model images without pre-processing. We constructed a dataset consisting of 50 greyscale 32x32 pixel images of two individuals from the UMIST faces dataset (Graham and Allinson, 1998). Both series of images show a person turning his head to the right to varying degrees.

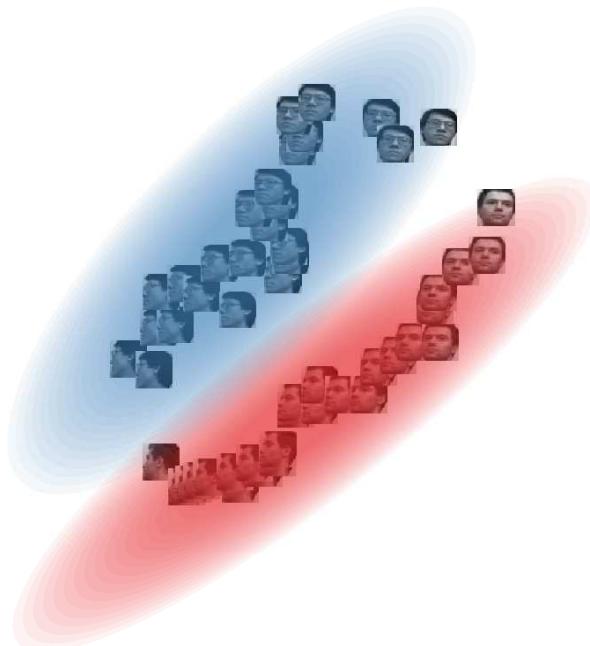


Figure 7.5 A sample from the 2-dimensional latent space when modeling a series of 32x32 face images. Images are rendered at their latent 2D coordinates. Our model correctly discovers that the data consists of two separate manifolds, both approximately one-dimensional, which both share the same head-turning structure.

Figure 7.5 shows a sample from the posterior over latent coordinates as well as the density model, with each image rendered at its location in the latent space. The model has recovered three relevant, interpretable features of the dataset: First, that there are two distinct faces. Second, that each set of images lies approximately along a smooth one-dimensional manifold. Third, that the two manifolds share roughly the same structure: the front-facing images of both individuals lie close to one another, as do the side-facing images.

### 7.5.3 Density estimation

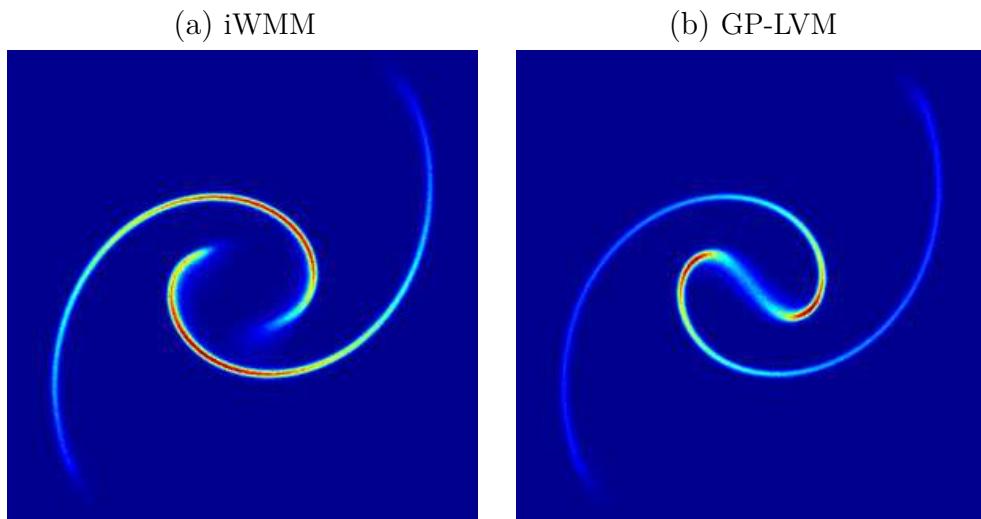


Figure 7.6 *Left*: Posterior density inferred by the iWMM in the observed space, on the 2-curve data. *Right*: Posterior density inferred by the iWMM with one component, a model equivalent to a fully-Bayesian GP-LVM.

Figure 7.6(a) shows the posterior density in the observed space inferred by the iWMM on the 2-curve data, computed using 1000 samples from the Markov chain. The two separate manifolds of high density implied by the two curved lines was recovered by the iWMM. Note also that the density along the manifold varies along with the density of data, shown in figure 7.4(a).

This result can be compared to a special case of our model with only a single Gaussian in the latent space, equivalent to a fully-Bayesian GP-LVM. Figure 7.6(b) shows that the single-cluster variant of the iWMM posterior is forced to place significant density connecting the two clusters, since it has to reproduce the observed density manifold by warping a single Gaussian.

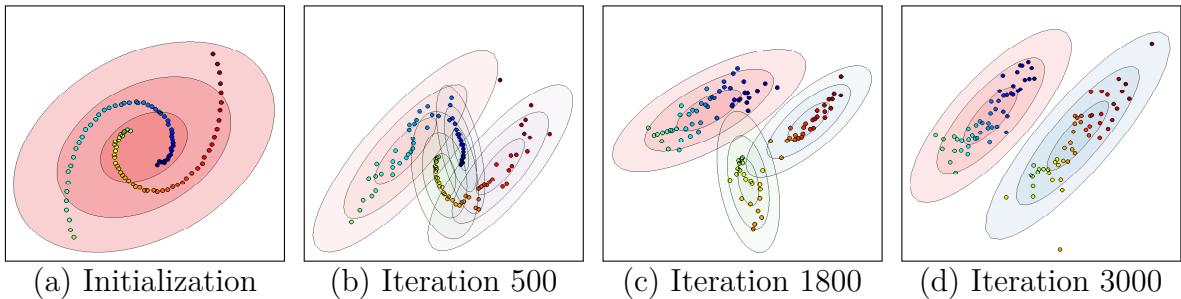


Figure 7.7 The inferred infinite GMMs over iterations in the two-dimensional latent space with the iWMM using the 2-curve data. Labels indicate the number of iterations of the sampler, and the color of each point represents its ordering in the observed coordinates.

#### 7.5.4 Mixing

An interesting side-effect of learning the number of latent clusters is that this added flexibility can help the sampler escape local minima. Figure 7.7 shows the samples of the latent coordinates and clusters of the iWMM over a single Markov chain modeling the 2-curve data. Figure 7.7(a) shows the latent coordinates initialized at the observed coordinates, starting with one latent component. After 500 iterations, each curved line was modeled by two components. After 1800 iterations, the left curved line was modeled by a single component. After 3000 iterations, the right curved line was also modeled by a single component, and the dataset was appropriately clustered. This configuration was relatively stable, and a similar state was found at the 5000th iteration.

#### 7.5.5 Visualization

Next, we briefly investigate the potential of the iWMM for low-dimensional visualization of data. Figure 7.8(a) shows the latent coordinates obtained by averaging over 1000 samples from the posterior of the iWMM. Because rotating the latent coordinates does not change their probability, simple averaging may not be an adequate way to summarize the posterior. However, we show this result in order to show the characteristics of latent coordinates obtained by the iWMM. The estimated latent coordinates are clearly separated, and they form two straight lines. This result is an example of the iWMM recovering the original topology of the data before it was warped.

For comparison, Figure 7.8(b) shows the latent coordinates estimated by the iWMM when forced to use a single cluster (again, equivalent to a fully-Bayesian GP-LVM). In this case, the latent coordinates lie in two sections of a single straight line.

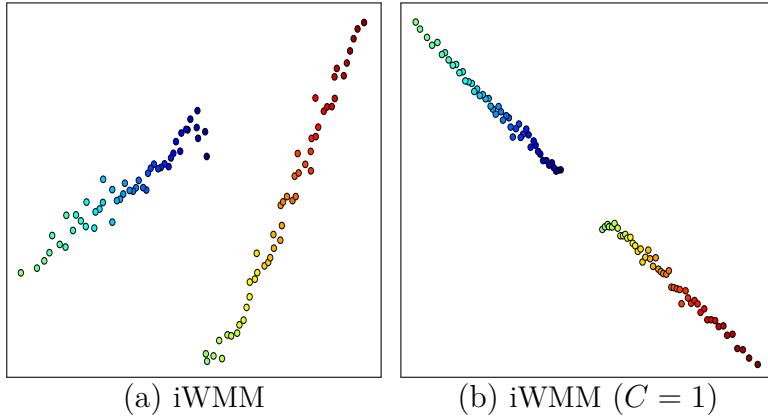


Figure 7.8 Latent coordinates of the 2-curve data, estimated by two different methods.

Regardless of the dimension of the latent space, the iWMM will tend to model each cluster with as low-dimensional a Gaussian as possible, since a narrowly-shaped Gaussian will assign the latent coordinates much higher likelihood than a spherical Gaussian.

### 7.5.6 Clustering performance

We more formally evaluated the density estimation and clustering performance of the proposed model using four real datasets: iris, glass, wine and vowel, obtained from LIBSVM multi-class datasets (Chang and Lin, 2011), in addition to the four synthetic datasets shown above: 2-curve, 3-semi, 2-circle and pinwheel (Adams and Ghahramani, 2009). The statistics of these datasets are summarized in Table 7.1. In each experiment,

Table 7.1 Statistics of the datasets used for evaluation.

	2-curve	3-semi	2-circle	pinwheel	iris	glass	wine	vowel
samples: $N$	100	300	100	250	150	214	178	528
dimension: $D$	2	2	2	2	4	9	13	10
num. clusters: $C$	2	3	2	5	3	7	3	11

we show the results of ten-fold cross-validation. Results in bold are not significantly different from the best performing method in each column according to a paired t-test.

Table 7.2 compares the clustering performance of the iWMM with the iGMM, quantified by the Rand index (Rand, 1971), which measures the correspondence between inferred clusters and true clusters. Since the manifold on which the observed data lies can be at most  $D$ -dimensional, we set the latent dimension  $Q$  equal to the observed

Table 7.2 Average Rand index for evaluating clustering performance.

	2-curve	3-semi	2-circle	Pinwheel	Iris	Glass	Wine	Vowel
iGMM	0.52	0.79	0.83	0.81	0.78	0.60	0.72	<b>0.76</b>
iWMM( $Q=2$ )	<b>0.86</b>	<b>0.99</b>	<b>0.89</b>	<b>0.94</b>	<b>0.81</b>	<b>0.65</b>	0.65	0.50
iWMM( $Q=D$ )	<b>0.86</b>	<b>0.99</b>	<b>0.89</b>	<b>0.94</b>	0.77	0.62	<b>0.77</b>	<b>0.76</b>

dimension  $D$  in iWMMs. We also included the  $Q = 2$  case in an attempt to characterize how much modeling power is lost by forcing the latent representation to be visualizable.

These experiments were designed to measure the extent to which nonparametric cluster shapes helped to estimate meaningful clusters. To eliminate any differences due to different inference procedures, we used identical code for the iGMM and iWMM, the only difference being that the warping function was set to the identity  $\mathbf{x} = \mathbf{y}$ . Both variants of the iWMM usually outperformed the iGMM on this measure.

### 7.5.7 Density estimation

Next, we compared the iWMM in terms of predictive density against kernel density estimation (KDE) and the (iGMM). For KDE, the kernel width was estimated by maximizing the leave-one-out density.

Although all these methods are consistent density estimators, we expect the iWMM to have an advantage for finite datasets, since its density contours can follow the data density. Table 7.3 lists average test log likelihoods.

Table 7.3 Average test log-likelihoods for evaluating density estimation performance.

	2-curve	3-semi	2-circle	Pinwheel	Iris	Glass	Wine	Vowel
KDE	-2.47	-0.38	-1.92	-1.47	<b>-1.87</b>	1.26	-2.73	<b>6.06</b>
iGMM	-3.28	-2.26	-2.21	-2.12	-1.91	3.00	<b>-1.87</b>	-0.67
iWMM( $Q=2$ )	<b>-0.90</b>	<b>-0.18</b>	<b>-1.02</b>	<b>-0.79</b>	<b>-1.88</b>	<b>5.76</b>	<b>-1.96</b>	<b>5.91</b>
iWMM( $Q=D$ )	<b>-0.90</b>	<b>-0.18</b>	<b>-1.02</b>	<b>-0.79</b>	<b>-1.71</b>	<b>5.70</b>	-3.14	-0.35

The iWMM usually achieved higher test likelihoods than the KDE and the iGMM. The sometimes large differences between performance in the  $D = 2$  case and the  $D = Q$  case may be attributed to the fact that when the latent dimension is high, it requires many samples from the latent distribution to produce an accurate estimate of the posterior density at the test locations. This difficulty in inference might be solved by using a warping with back-constraints (Lawrence, 2006), that would allow a more direct evaluation of the density at a given point in the observed space.

### Source code

Code to reproduce all the above figures and experiments is available at [github.com/dvvenaud/warped-mixtures](https://github.com/dvvenaud/warped-mixtures).

## 7.6 Conclusions

This chapter introduced a simple generative model of non-Gaussian density manifolds which can infer nonlinearly separable clusters, low-dimensional representations of varying dimension per cluster, and density estimates which smoothly follow the contours of each cluster. We then introduced a sampler for this model which integrates out both the cluster parameters and the warping function exactly.

Non-probabilistic methods such as spectral clustering can also produce nonparametric cluster shapes, but usually lack principled methods for setting kernel parameters, the number of clusters, and the implicit dimension of the learned manifolds, other than by cross-validation. This chapter shows that using a fully generative model allows most model choices to be determined automatically.

Many methods have been proposed which can perform some combination of clustering, manifold learning, density estimation and visualization. We demonstrated that a simple but flexible probabilistic generative model can perform well at all these tasks.

## 7.7 Future work

### More sophisticated latent density models

The Dirichlet process mixture of Gaussians in the latent space of our model could easily be replaced by a more sophisticated density model, such as a hierarchical Dirichlet process (Teh et al., 2006), or a Dirichlet diffusion tree (Neal, 2003). Another straightforward extension of our model would be making inference more scalable by using sparse Gaussian processes (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006) or more advanced Hamiltonian Monte Carlo methods (Zhang and Sutton, 2011).

### A finite cluster count model

Miller and Harrison (2013) note that the Dirichlet process assumes infinitely many clusters, and that estimates of the number of clusters in a dataset based on Bayesian inference are inconsistent under this model. They propose a consistent alternative which

also allows efficient Gibbs sampling, called the mixture of finite mixtures. Replacing the Dirichlet process with a mixture of finite mixtures would improve the consistency properties of the iWMM in most applications.

### Semi-supervised learning

An interesting but more complex extension of the iWMM would be a semi-supervised version of the model. The iWMM could allow label propagation along regions of high density in the latent space, even if the individual points in those regions are stretched far apart along low-dimensional manifolds in the observed space. Another natural extension would be to allow a separate warping for each cluster, producing a mixture of warped Gaussians, rather than a warped mixture of Gaussians.

### Learning the topology of data manifolds

Some datasets naturally live on manifolds which are not simply connected. For example, motion capture data or video of a person walking in a circle naturally lives on a torus, with one coordinate specifying the phase of the person's step, and another specifying how far around the circle they've walked.

As shown in section 2.8, using structured kernels to specify the warping of a latent space gives rise to interesting topologies on the observed density manifold. If a suitable method for computing the marginal likelihood of a GP-LVM is available, an automatic search similar to that described in chapter 3 would be possible, automatically finding the topology of the data manifold.

# Chapter 8

## Bayesian Estimation of Integrals

### 8.1 Log-Gaussian Process Models

When integrating

$$Z = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.1)$$

$$= \mathbb{E}_{p(\mathbf{x})} [f(\mathbf{x})] \quad (8.2)$$

We typically model the function  $f(\mathbf{x})$  with a Gaussian process prior. However, for many problems in Bayesian inference,  $f(\mathbf{x})$  is a likelihood function, typically of the form  $f(\mathbf{x}) = \exp(\log \ell(\mathbf{x}))$ . This implies that  $f(\mathbf{x})$  is both positive and that it has a high dynamic range, making a stationary GP a poor model for this function.

A much more believable model would be a GP prior on  $\log \ell(\mathbf{x})$ . A GP prior on  $\log \ell$  induces a log GP prior on  $\ell(\mathbf{x})$ . However, this model makes the expectation of the integral over  $f$  seemingly intractable:

$$\mathbb{E}[Z] = \mathbb{E}_{\text{GP}(\log \ell)} \left[ \int \exp(\log \ell(\mathbf{x})) p(\mathbf{x}) d\mathbf{x} \right] \quad (8.3)$$

$$= \int \int \exp(\log \ell(\mathbf{x})) p(\mathbf{x}) p(\log \ell) d\mathbf{x} d\log \ell \quad (8.4)$$

$$= \int \int \exp(\log \ell(\mathbf{x})) p(\log \ell(\mathbf{x})) d\log \ell(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.5)$$

$$= \int \mathbb{E}_{\text{GP}(\log \ell)} [\exp(\log \ell(\mathbf{x}))] p(\mathbf{x}) d\mathbf{x} \quad (8.6)$$

Where the posterior mean function of  $\exp(\log \ell)$ ) is:

$$\mathbb{E}_{\text{GP}(\log \ell)} [\exp(\log \ell(\mathbf{x}))] = \exp \left[ \mathbb{E}_{\text{GP}(\log \ell)} [\log \ell(\mathbf{x})] + \frac{1}{2} \mathbb{V}_{\text{GP}(\log \ell)} [\log \ell(\mathbf{x})] \right] \quad (8.7)$$

$$= \exp \left[ k(\mathbf{x}, \mathbf{X}) K^{-1} \mathbf{y} + \frac{1}{2} (k(x, x) - k(\mathbf{x}, \mathbf{X}) K^{-1} k(\mathbf{X}, \mathbf{x})) \right] \quad (8.8)$$

$$= m_{\exp}(\mathbf{x}) \quad (8.9)$$

Which is not known to have a tractable form for common choices of covariance functions.

However, we can approximate this integral by fitting a GP to points  $\mathbf{x}_c$  along the function  $m_{\exp}(\mathbf{x})$ , and exactly integrating under that integral. We will call the mean of this GP  $\bar{\ell}(\mathbf{x})$ . We can then compute a mean estimate,  $Z_0$ , by exactly integrating under this GP:

$$Z_0 = \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.10)$$

$$= z^T K^{-1} \left[ \exp \left( m_{\log \ell}(\mathbf{x}_c) + \frac{1}{2} V_{\log \ell}(\mathbf{x}_c) \right) \right] \quad (8.11)$$

Where  $z$  is defined below.

The approximation  $Z_0$  exactly approaches  $\mathbb{E}_{\text{GP}(\log \ell)}[Z]$  in the limit of infinite candidate points  $\mathbf{x}_c$ . This is important, since we believe that a GP is a much better model of  $\log \ell(\mathbf{x})$  than of  $\ell(\mathbf{x})$ .

The mean function  $\bar{\ell}$  will be inflated in regions of high marginal variance in  $\log \ell$ . If we were simply to fit a GP to  $\exp(m_{\log \ell}(\mathbf{x}_c))$ , we would underestimate  $\mathbb{E}_{\text{GP}(\log \ell)}[Z]$  even in the limit of infinite samples.

## 8.2 Linearization

Note that  $Z_0$  is a functional of  $\log \ell$ :

$$Z[\log \ell] = \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.12)$$

Here we approximately compute  $\mathbb{E}[Z]$  when we have only a sparse set of samples. We will first linearize  $Z$  as a functional of  $\log \ell$  around the posterior mean,  $\log \ell_0$ .

$$Z[\log \ell] = \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.13)$$

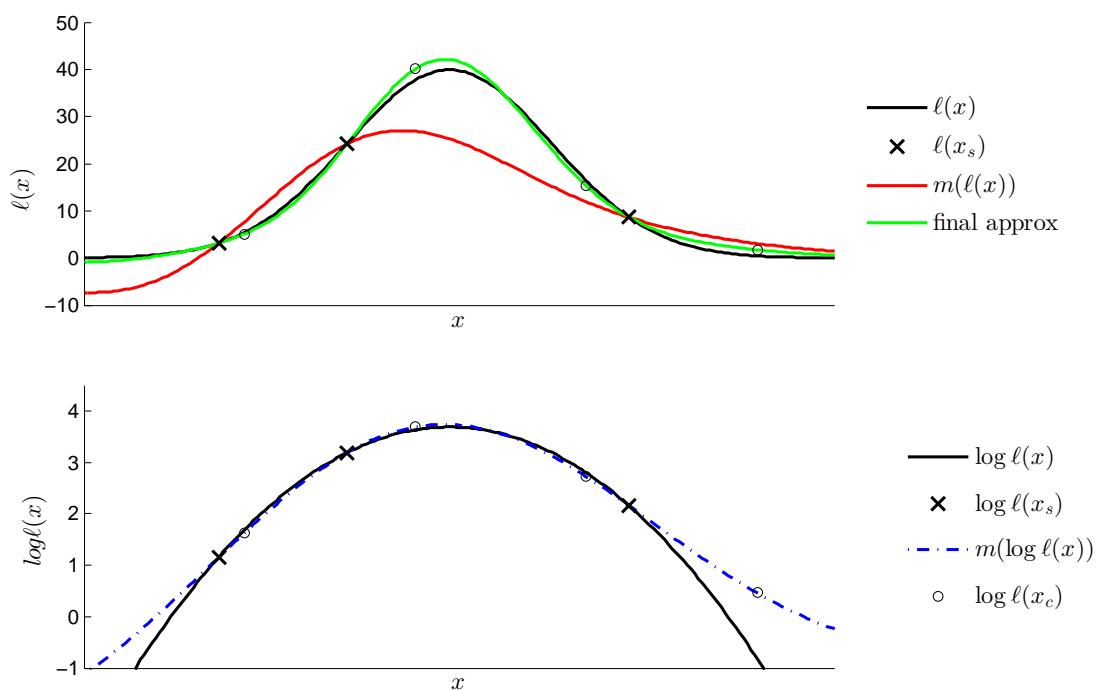


Figure 8.1 A GP versus a log-GP

$$\log \ell_0(\mathbf{x}) = \mathbb{E}_{\text{GP}(\log \ell)} [\log \ell(\mathbf{x})] \quad (8.14)$$

$$\hat{Z}[\log \ell] := Z_0 + \varepsilon[\log \ell] \quad (8.15)$$

$$Z_0 = \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.16)$$

$$\varepsilon[\log \ell] = \int \left( \frac{\partial \hat{Z}}{\partial \log \ell(\mathbf{x})} \Big|_{\log \ell_0(\mathbf{x})} \right) [\log \ell(\mathbf{x}) - \log \ell_0(\mathbf{x})] d\mathbf{x} \quad (8.17)$$

We do not have a closed-form expression for  $\frac{\partial \hat{Z}}{\partial \log \ell(\mathbf{x})} \Big|_{\log \ell_0(\mathbf{x})}$  everywhere, but at the points  $\mathbf{x}_c$ , it will be

$$\frac{\partial \hat{Z}}{\partial \log \ell(\mathbf{x}_c)} \Big|_{\log \ell_0(\mathbf{x}_c)} = \frac{\partial \hat{Z}}{\partial \log \ell(\mathbf{x}_c)} \Big|_{\log \ell_0(\mathbf{x}_c)} \quad (8.18)$$

$$= \bar{\ell}(\mathbf{x}) p(\mathbf{x}) \quad (8.19)$$

We hope that  $\bar{\ell}(\mathbf{x})$  is a good approximation everywhere else, and make the approximation:

$$\frac{\partial Z}{\partial \log \ell(\mathbf{x})} \Big|_{\log \ell_0(\mathbf{x})} \approx \bar{\ell}(\mathbf{x}) p(\mathbf{x}) \quad (8.20)$$

so

$$\varepsilon[\log \ell] \approx \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) [\log \ell(\mathbf{x}) - \log \ell_0(\mathbf{x})] d\mathbf{x} \quad (8.21)$$

Which is the same as in the SBQ paper, except that  $Z_0$  will be higher than it would be in the SBQ setup, because  $\bar{\ell}$  will be inflated due to uncertainty in  $\log \ell$ .

### 8.2.1 Mean of $\hat{Z}$

Under this linearization, the mean of  $\hat{Z}$  is simply

$$\mathbb{E}_{\text{GP}(\log \ell)}[\hat{Z}] = \mathbb{E}_{\text{GP}(\log \ell)}[Z_0 + \varepsilon[\log \ell]] \quad (8.22)$$

$$= Z_0 \quad (8.23)$$

Again,  $Z_0$  will be higher than it would be in the SBQ setup, because  $\bar{\ell}$  will be inflated due to uncertainty in  $\log \ell$ .

Now we derive a closed-form expression for the case where the prior is  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|b, B)$ , the kernel function  $k_\ell(x_\star, \mathbf{x}') = h_\ell \mathcal{N}(\mathbf{x}|\mathbf{x}', A_\ell)$ , and the mean function  $\mu_\ell(\mathbf{x}) = \mu_\ell$ .

$$Z_0 = \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.24)$$

$$= \mathbb{E}_{\mathcal{GP}(\ell(\mathbf{x}))} \left[ \int \ell(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \right] \quad (8.25)$$

$$= \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.26)$$

where

$$\bar{\ell}(\mathbf{x}_\star) = \mu(\mathbf{x}_\star) + k_\ell(\mathbf{x}_\star, \mathbf{x}_c) K_\ell(\mathbf{x}_c, \mathbf{x}_c)^{-1} (m_{\text{exp}}(\mathbf{x}_c) - \mu(\mathbf{x}_c)) \quad (8.27)$$

$$= \mu(\mathbf{x}_\star) + \sum_{i \in c} k_\ell(\mathbf{x}_\star, \mathbf{x}_i) \beta_i \quad (8.28)$$

$$= \mu(\mathbf{x}_\star) + \sum_{i \in c} h_\ell \mathcal{N}(\mathbf{x}_\star | \mathbf{x}_i, A_\ell) \beta_i \quad (8.29)$$

where

$$\beta_i = K_\ell(\mathbf{x}_c, \mathbf{x}_c)^{-1} (m_{\text{exp}}(\mathbf{x}_i) - \mu(\mathbf{x}_i)) \quad (8.30)$$

$$(8.31)$$

are simply constants. So

$$Z_0 = \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (8.32)$$

$$Z_0 = \int \left[ \mu_\ell + \sum_{i \in c} h_\ell \mathcal{N}(\mathbf{x} | \mathbf{x}_i, \mathbf{A}_\ell) \beta_i \right] \mathcal{N}(\mathbf{x} | \mathbf{b}, \mathbf{B}) d\mathbf{x} \quad (8.33)$$

$$Z_0 = \mu_\ell + h_\ell \int \left[ \sum_{i \in c} \mathcal{N}(\mathbf{x} | \mathbf{x}_i, \mathbf{A}_\ell) \beta_i \right] \mathcal{N}(\mathbf{x} | \mathbf{b}, \mathbf{B}) d\mathbf{x} \quad (8.34)$$

$$Z_0 = \mu_\ell + h_\ell \sum_{i \in c} \beta_i \int \mathcal{N}(\mathbf{x} | \mathbf{x}_i, \mathbf{A}_\ell) \mathcal{N}(\mathbf{x} | \mathbf{b}, \mathbf{B}) d\mathbf{x} \quad (8.35)$$

and the integral

$$z_i = \int \mathcal{N}(\mathbf{x} | \mathbf{x}_i, \mathbf{A}_\ell) \mathcal{N}(\mathbf{x} | \mathbf{b}, \mathbf{B}) d\mathbf{x} \quad (8.36)$$

$$= \text{ProdNormZ}(\mathbf{x}_i, \mathbf{A}_\ell, \mathbf{b}, \mathbf{B}) \quad (8.37)$$

$$(8.38)$$

where

$$\text{ProdNormZ}(a, A, b, B) = \mathcal{N}(\mathbf{x} | B(A+B)^{-1}a + A(A+B)^{-1}b, A(A+B)^{-1}B) \mathcal{N}(a | b, A+B) \quad (8.39)$$

### 8.2.2 Variance of $\hat{Z}$

$$\mathbb{V}[\hat{Z}] = \mathbb{E}[\hat{Z}^2] - (\mathbb{E}[\hat{Z}])^2 \quad (8.40)$$

$$\mathbb{V}[\hat{Z}] = \mathbb{E}[\hat{Z}^2] - Z_0^2 \quad (8.41)$$

$$= \int (\hat{Z}[\log \ell])^2 p(\log \ell) d\log \ell - Z_0^2 \quad (8.42)$$

$$= \int [Z_0 + \varepsilon(\log \ell)]^2 p(\log \ell) d\log \ell - Z_0^2 \quad (8.43)$$

$$= \int [Z_0^2 + 2Z_0\varepsilon(\log \ell) + (\varepsilon(\log \ell))^2] p(\log \ell) d\log \ell - Z_0^2 \quad (8.44)$$

$$= \int [2Z_0\varepsilon(\log \ell) + (\varepsilon(\log \ell))^2] p(\log \ell) d\log \ell \quad (8.45)$$

$$= \int 2Z_0\varepsilon(\log \ell) p(\log \ell) d\log \ell + \int (\varepsilon(\log \ell))^2 p(\log \ell) d\log \ell \quad (8.46)$$

$$= 0 + \int (\varepsilon(\log \ell))^2 p(\log \ell) d\log \ell \quad (8.47)$$

$$= \int \left( \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) [\log \ell(\mathbf{x}) - \log \ell_0(\mathbf{x})] d\mathbf{x} \right)^2 p(\log \ell) d\log \ell \quad (8.48)$$

$$= \int \left[ \int \bar{\ell}(\mathbf{x}) p(\mathbf{x}) [\log \ell(\mathbf{x}) - \log \ell_0(\mathbf{x})] d\mathbf{x} \right] \left[ \int \bar{\ell}(\mathbf{x}') p(\mathbf{x}') [\log \ell(\mathbf{x}') - \log \ell_0(\mathbf{x}')] d\mathbf{x}' \right] p(\log \ell) d\log \ell$$

$$= \iiint \left( [\log \ell(\mathbf{x}) - \log \ell_0(\mathbf{x})] [\log \ell(\mathbf{x}') - \log \ell_0(\mathbf{x}')] p(\log \ell) d\log \ell \right) \bar{\ell}(\mathbf{x}) \bar{\ell}(\mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}'$$

$$= \iint \mathbb{E}_{p(\log \ell)} \left[ \log \ell(\mathbf{x}) - \mathbb{E} [\log \ell(\mathbf{x})] \right] \left[ \log \ell(\mathbf{x}') - \mathbb{E} [\log \ell(\mathbf{x}')] \right] \bar{\ell}(\mathbf{x}) \bar{\ell}(\mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}'$$

$$= \iint \text{Cov}_{\log \ell}(\mathbf{x}, \mathbf{x}') \bar{\ell}(\mathbf{x}) \bar{\ell}(\mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \quad (8.49)$$

Which has a nice form: It is simply the standard BMC variance estimate, but scaled to be proportional to the height of  $\bar{\ell}(\mathbf{x})$ .

If we assume that  $k_{\log \ell}(x_\star, \mathbf{x}') = h_{\log \ell} \mathcal{N}(\mathbf{x}_\star | \mathbf{x}', A_{\log \ell})$ , we can continue the derivation to get a closed-form solution:

$$\text{Cov}_{\log \ell}(\mathbf{x}_\star, \mathbf{x}'_\star) = k(\mathbf{x}_\star, \mathbf{x}') - k(\mathbf{x}_\star, \mathbf{x}_s) K(\mathbf{x}_s, \mathbf{x}_s)^{-1} k(\mathbf{x}_s, \mathbf{x}'_\star) \quad (8.50)$$

$$= k(\mathbf{x}_\star, \mathbf{x}') - k(\mathbf{x}_\star, \mathbf{x}_s) K_{\log \ell}^{-1} k(\mathbf{x}_s, \mathbf{x}'_\star) \quad (8.51)$$

so

$$\mathbb{V}[\hat{Z}] = \iint \text{Cov}_{\log \ell}(\mathbf{x}, \mathbf{x}') \bar{\ell}(\mathbf{x}) \bar{\ell}(\mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \quad (8.52)$$

$$= \iint [k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{x}_s) K_{\log \ell}^{-1} k(\mathbf{x}_s, \mathbf{x}')] \quad (8.53)$$

$$\times \left[ \mu(\mathbf{x}) + \sum_{i \in c} k_\ell(\mathbf{x}, \mathbf{x}_i) \beta_i \right] \left[ \mu(\mathbf{x}') + \sum_{i' \in c} k_\ell(\mathbf{x}', \mathbf{x}_{i'}) \beta_{i'} \right] p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \quad (8.54)$$

$$= \iint [h_{\log \ell} \mathcal{N}(\mathbf{x} | \mathbf{x}', A_{\log \ell}) - h_{\log \ell}^2 \mathcal{N}(\mathbf{x} | \mathbf{x}_s, A_{\log \ell}) K_{\log \ell}^{-1} \mathcal{N}(\mathbf{x}' | \mathbf{x}_s, A_{\log \ell})] \quad (8.55)$$

$$\times \left[ \mu_\ell + h_\ell \sum_{i \in c} \mathcal{N}(\mathbf{x} | \mathbf{x}_i, A_\ell) \beta_i \right] \left[ \mu_\ell + h_\ell \sum_{i' \in c} \mathcal{N}(\mathbf{x}' | \mathbf{x}_{i'}, A_\ell) \beta_{i'} \right] \quad (8.56)$$

$$\times \mathcal{N}(\mathbf{x} | \mathbf{b}, \mathbf{B}) \mathcal{N}(\mathbf{x}' | \mathbf{b}, \mathbf{B}) d\mathbf{x} d\mathbf{x}' \quad (8.57)$$

### 8.2.3 Marginal Variance

The marginal variance of  $\bar{\ell}(\mathbf{x})$  when integrating over  $\log \ell(\mathbf{x})$  can be found by evaluating (8.58) at a single point:

$$\mathbb{V}[\bar{\ell}(\mathbf{x})] = \mathbb{V}_{\log \ell}(\mathbf{x}) \bar{\ell}(\mathbf{x})^2 \quad (8.58)$$

which is again proportional to the height of  $\bar{\ell}(\mathbf{x})$ , a desirable property that can now be handled in closed form.

## 8.3 Nonparametric Inference via Bayesian Quadrature

We extend the approximate integration method of Bayesian Quadrature to infinite structured domains, introducing a new inference method for nonparametric models. This method admits more flexible sampling schemes than Markov chains, for example active learning, and provides natural convergence diagnostics. We give conditions necessary for consistency, and show how to construct kernels between structures which take advantage of symmetries in the likelihood function. We demonstrate our inference method on both the Dirichlet process mixture model and the Indian buffet process.

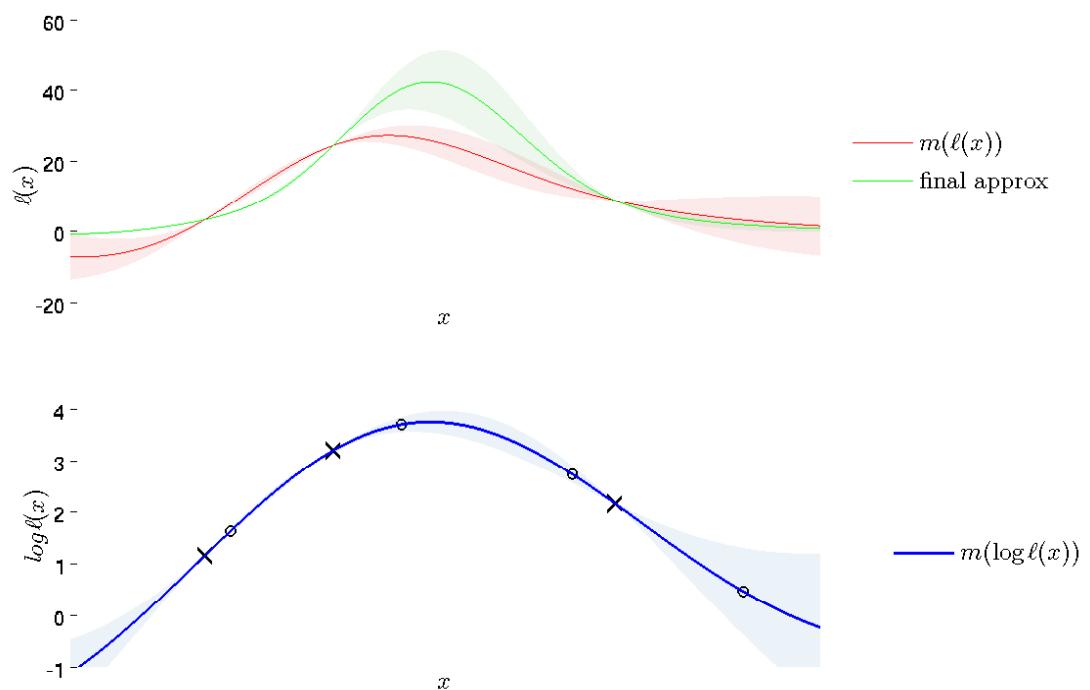


Figure 8.2 A GP versus a log-GP

## 8.4 Introduction

The central problem of probabilistic inference is to compute integrals over probability distributions of the form

$$Z_{f,p} = \int f(\theta)p(\theta)d\theta \quad (8.59)$$

Examples include computing marginal distributions, making predictions while marginalizing over parameters, or computing the Bayes risk in a decision problem. Machine learning has produced a rich set of methods for computing these integrals, such as Expectation Propagation[cite], Variational Bayes[cite], and many variations of Markov chain Monte Carlo (MCMC ) [cite Iain Murray].

In non-parametric models, estimating (8.59) is especially challenging, as the domain of integration in is infinite-dimensional. A variety of MCMC methods have been developed to tackle this problem. However, MCMC has known weaknesses, such as difficulty diagnosing convergence, the requirement of a burn-in period, and difficulty obtaining samples from a given subset of possible  $\theta$ .

Bayesian quadrature (BQ) ?, also known as Bayesian Monte Carlo ?, is a model-based method of approximate integration. BQ infers a posterior distribution over  $f$  conditioned on a set of samples  $f(\boldsymbol{\theta}_s)$ , and gives a posterior distribution on  $Z_{f,p}$ . BQ remains a relatively unexplored integration method, and has so far only been used in low-dimensional, continuous spaces ?.

**Summary of contributions** In this paper, we extend the BQ method to infinite, structured domains, introducing a new family of inference methods for non-parametric models. We give conditions necessary for consistency. We introduce kernels for inference problems using the Indian Buffet process and Dirichlet Proces mixture model which encode the many symmetries of the likelihood functions.

We then demonstrate the advantages of model-based inference on synthetic datasets, such as uncertainty estimates, flexibility in sampling methods, and post-hoc sensitivity analysis of prior and likelihood parameters. We then discuss limitations of the framework as it stands.

## 8.5 Bayesian Quadrature

In contrast to MCMC, a procedure which computes (8.59) in the limit of infinite samples, Bayesian quadrature is a *model-based* integration method. This means that BQ puts a

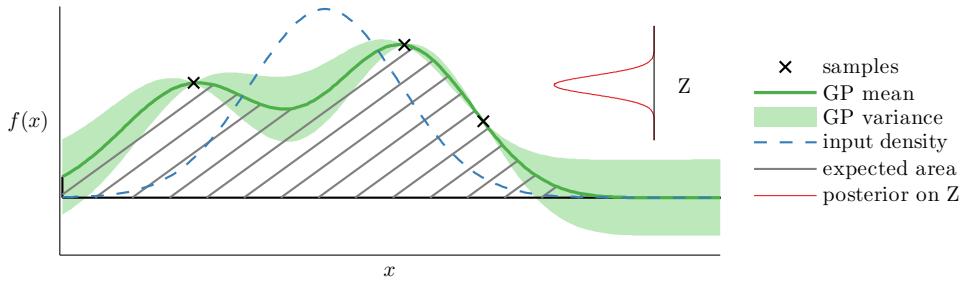


Figure 8.3 An illustration of Bayesian quadrature. The function  $f(x)$  is sampled at a set of input locations. This induces a Gaussian process posterior distribution on  $f$ , which is integrated in closed form against the target density,  $p(\mathbf{x})$ . Since the amount of volume under  $f$  is uncertain, this gives rise to a (Gaussian) posterior distribution over  $Z_{f,p}$ .

prior distribution on  $f$ , then conditions on some observations  $f(\boldsymbol{\theta}_s)$  at some query points  $\boldsymbol{\theta}_s$ . The posterior distribution over  $f$  then implies a distribution over  $Z_{f,p}$ , which can be obtained by integrating over all possible  $f$ . See Figure 8.3 for an illustration of Bayesian Quadrature.

It may seem circular to introduce an integral over an uncountable number of functions in order to solve what was originally an integral over a single function. However, the GP posterior has a simple form which makes integration possible in closed form in many cases. If  $f$  is assigned a Gaussian process prior with kernel function  $k$  and mean 0, then after conditioning on function evaluations  $\mathbf{y} = f(\boldsymbol{\theta}_1) \dots f(\boldsymbol{\theta}_N)$ , we obtain:

$$p(\mathbf{f}(\boldsymbol{\theta}_*) | \mathbf{y}) = \mathcal{N}\left(\mathbf{f}(\boldsymbol{\theta}_*) \mid \bar{\mathbf{f}}(\boldsymbol{\theta}_*), \text{cov}(\boldsymbol{\theta}_*, \boldsymbol{\theta}'_*)\right) \quad (8.60)$$

where

$$\bar{\mathbf{f}}(\boldsymbol{\theta}_*) = k(\boldsymbol{\theta}_*, \boldsymbol{\theta}_s) K^{-1} \mathbf{y} \quad (8.61)$$

$$\text{cov}(\mathbf{x}_*, \mathbf{x}'_*) = k(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) - k(\boldsymbol{\theta}_*, \boldsymbol{\theta}_s) \mathbf{K}^{-1} k(\boldsymbol{\theta}_s, \boldsymbol{\theta}_*) \quad (8.62)$$

and  $\mathbf{K}_{mn} = k(\boldsymbol{\theta}_m, \boldsymbol{\theta}_n)$ . Conveniently, the GP posterior implies a closed form for the expectation and variance of (8.59):

$$\mathbb{E}[Z_{f,p} | \mathbf{y}] = \mathbb{E}_{\text{GP}(f|\mathbf{y})} \left[ \int f(\theta) p(\theta) d\theta \right] = \left[ \int k(\theta, \boldsymbol{\theta}_s) p(\theta) d\theta \right] \mathbf{K}^{-1} \mathbf{y} = \mathbf{z}^\top \mathbf{K}^{-1} \mathbf{y} \quad (8.63)$$

$$\mathbb{V}[Z_{f,p} | \mathbf{y}] = \mathbb{V}_{\text{prior}}[Z_{f,p}] - \mathbf{z}^\top \mathbf{K}^{-1} \mathbf{z} \quad (8.64)$$

where

$$z_n = \int k(\theta, \theta_n) p(\theta) d\theta \quad (8.65)$$

$$\mathbb{V}_{\text{prior}} [Z_{f,p}] = \iint k(\theta, \theta') p(\theta) p(\theta') d\theta d\theta'. \quad (8.66)$$

For longer derivations, see the supplementary material. This brings us to the one of the main technical constraints of this method: Choosing a form for the kernel function  $k(\theta, \theta')$  such that we can compute (8.65) and (8.66) efficiently. We must also set or integrate out any parameters of  $k$ .

### 8.5.1 BQ as an inference method

If we assume that in (8.59), the form of  $p(\theta)$  is such that we can compute  $z_n$ , then applying BQ is straightforward. For example, in ?, BQ was used to solve problems where  $p(\theta)$  was a Gaussian prior over parameters, and  $f(\theta) = p(x|\theta)$  was a likelihood function, making  $Z$  the *model evidence*, a useful quantity when comparing models.

Typically, however, we are also interested in other quantities besides the model evidence, such as marginal distributions of latent parameter. In that case, we must solve a more difficult integral, where  $p(\theta) = p^*(\theta|x)$  is a possibly unnormalized distribution of unknown form.

$$\mathbb{E}_{p(\theta|x)} [f(\theta)] = \int f(\theta) p(\theta|x) d\theta = \frac{1}{Z} \int f(\theta) p(x|\theta) p(\theta) d\theta \quad \text{where } Z = \int p(x|\theta) p(\theta) d\theta \quad (8.67)$$

Integrals of this form can also be solved using Bayesian quadrature. For a thorough treatment of this problem, see [Mike's BQR paper].

This will show that BQ can be applied in several ways: For instance, if we wish to perform inference about an unknown parameter  $\tau$ , we can include it as a variable to be integrated over by the GP. However, if for some technical reason this is difficult, we can also simply vary our extra parameter over some range, recomputing  $\mathbf{y}(\tau)$  and obtaining a marginal distribution over  $Z_{f,p}$  for each value of  $\tau$ . [Reference an experiment?] This approach is useful for post-hoc sensitivity analysis.

## 8.6 Guidelines for Constructing a Kernel

As pointed out above, one of the most important design decisions in BQ is the choice of kernel function  $k(\theta, \theta')$ , which specifies the prior covariance between the values of the likelihood function  $p(\mathbf{x}|\theta)$  and  $p(\mathbf{x}|\theta')$ . This function is somewhat analogous to the proposal distribution required to construct a Metropolis-Hastings sampler [cite]. In this section, we give guidance on how to construct an appropriate kernel.

The kernel typically should encode as much prior knowledge about the function being modeled as possible. In regression problems, this usually amounts to specifying the smoothness properties of the function being modeled. When doing inference, however, we typically know the likelihood function in closed form. The more properties of the likelihood function we can encode in the kernel, the fewer samples we will need in order to learn about the value of its integral. In particular, we should encode any known symmetries:

**Symmetry Encoding:** The prior correlation  $\frac{k(\theta, \theta')}{\sqrt{k(\theta, \theta)}\sqrt{k(\theta', \theta')}}$  should equal 1 when  $f(\theta) = f(\theta')$ . That is to say, if two parameter settings are indistinguishable under the likelihood, our model can converge more quickly if it enforces that those likelihood values are identical. This is another way of saying that the covariance function should encode all known symmetries.

The ability to encode symmetries in the kernel is one of the major advantages of BQ over Monte Carlo methods. In unidentifiable models such as mixture models, often it is known that there exist many symmetric modes in the posterior, which represents a major difficulty when computing model evidence. By encoding these symmetries in the prior over likelihood functions, BQ neatly solves the problem of identifiability when estimating  $Z_{f,p}$ .

### 8.6.1 Convergence

In the existing BQ literature [cite only 4 papers], the integration problems considered have been low-dimensional, and the kernel function used has always been, to the best of the authors' knowledge, the squared-exponential kernel. In that case, existing results on the consistency of GP regression [cite consistency] imply that, under some conditions, the BQ estimator (a linear transform of the GP posterior) is also consistent.

For infinite-dimensional spaces with complex kernels, it is more difficult to prove consistency, although the known structure of the likelihood functions may help. In this

section, we give conditions necessary for convergence.

First, the kernel must be positive-definite [Rasmussen and Williams \(2006\)](#). In addition, in order to ensure convergence, we must have that the following condition holds:

**Identifiability:** The prior correlation  $\frac{k(\theta, \theta')}{\sqrt{k(\theta, \theta)}\sqrt{k(\theta', \theta')}}$  must be less than 1 if it is possible that  $f(\theta) \neq f(\theta')$ . That is to say, if two values of the likelihood function can be different, our model must not enforce that those two function values are identical.

**Proposition 1.** *The above condition is necessary to guarantee convergence of the BQ posterior to the true value of  $Z$ .*

*Proof sketch.* Consider a prior  $p(\theta) = \frac{1}{2}\delta_{\theta_1}(\theta) + \frac{1}{2}\delta_{\theta_2}(\theta)$  where  $\theta_1 \neq \theta_2$ . If the prior correlation between  $f(\theta_1)$  and  $f(\theta_2)$  is one, then after observing one of those two values, say  $f(\theta_1)$  we have that  $\mathbb{E}_{\text{GP}}[Z] = f(\theta_1)$  and  $\mathbb{V}_{\text{GP}}[Z] = 0$ . However if  $f(\theta_1) \neq f(\theta_2)$ , then  $Z_0 = \frac{1}{2}f(\theta_1) + \frac{1}{2}f(\theta_2) \neq f(\theta_1)$ . Thus the estimator will have converged to the wrong hypothesis.  $\square$

For a more detailed proof, see the supplementary material. The statements in this section also hold for the problem of GP regression in general, but are specially relevant for the problem of inferring likelihood functions over latent structures. This is for two reasons: First, likelihood functions over structures can typically be shown to have many symmetries. Secondly, in the quadrature setting, we must learn about the function everywhere under the prior, not just in a small region or manifold, as is typical for regression problems. Exploiting the symmetries of the likelihood function is both possible, and necessary for fast convergence.

## 8.7 Inference in the Indian Buffet Process

In this section, we construct a kernel and demonstrate the use of BQ for inference in an infinite latent model, the Indian buffet process. The Indian buffet process (IBP) [?](#) is a distribution over binary matrices, usually used as a prior over latent features of a set of items. The IBP is nonparametric in the sense that the number of latent features is unbounded. For example, in [?](#), a model of images is constructed where the entries of a binary matrix specify which objects appear in which images. Although the number of objects is unknown beforehand, given a set of images, the flexible IBP prior allows inference on both the number of objects, and their presence over the dataset.

Under an IBP prior, the probability of seeing a matrix  $\mathbf{Z}$  with  $K$  columns is

$$P(\mathbf{Z}|\alpha) = \prod_{k=1}^K \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})} \quad (8.68)$$

where  $m_k$  is the number of ones in column  $k$ , and  $\alpha$  is the concentration parameter. There is an unfortunate clash of notation here, where  $Z_{f,p}$  denotes the model evidence, and  $\mathbf{Z}$  is used to denote a binary matrix.

To fully specify a model, we must also specify the likelihood of a set of observations  $\mathbf{X}$  given the latent structure,  $p(\mathbf{X}|\mathbf{Z})$ . For simplicity, we will use as a simple example the linear-Gaussian model from ?. This model assumes the data is generated as  $\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\epsilon}$ , with  $A_{ij} \sim \mathcal{N}(0, \sigma_A)$  and  $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_X)$ . The features in  $\mathbf{A}$  are turned on and off for each row of  $\mathbf{X}$  by the entries of  $\mathbf{Z}$ . Conveniently, we can integrate out the matrix  $\mathbf{A}$  to obtain a collapsed likelihood:

$$p(\mathbf{X}|\mathbf{Z}, \sigma_X, \sigma_A) = \frac{\exp \left\{ -\frac{1}{2\sigma_X^2} \text{tr}(\mathbf{X}^T(\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2}\mathbf{E}_J^{-1}\mathbf{Z}^T)\mathbf{X}) \right\}}{(2\pi)^{ND/2} \sigma_X^{(N-K)D} \sigma_A^{KD} |\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2}\mathbf{I}|^{\frac{D}{2}}} \quad (8.69)$$

The goal of inference in the IBP is usually do discover statistics about the matrices  $\mathbf{Z}$  and  $\mathbf{A}$ , or to produce a predictive distribution over new rows of  $\mathbf{X}$ . In this paper, we will show how to [...]

### 8.7.1 A kernel between binary matrices

Here we give a kernel which satisfies the guidelines given in section 8.6. Since the likelihood in (8.69) does not depend on the order of columns of  $\mathbf{Z}$ , we will construct a kernel function  $k(\mathbf{Z}, \mathbf{Z}')$  which is also invariant to permutations over columns of both  $\mathbf{Z}$  and  $\mathbf{Z}'$ :

$$k(\mathbf{Z}, \mathbf{Z}') = \sum_{k=1}^K \sum_{k'=1}^{K'} \sum_{n=1}^N \mathbf{Z}_{n,k} \mathbf{Z}'_{n,k'} \quad (8.70)$$

This kernel has the property that, for a given number of ones in  $\mathbf{Z}$  and  $\mathbf{Z}'$ , it attains its maximum value when every element of some permutation of columns of  $\mathbf{Z}$  is equal to  $\mathbf{Z}'$  (i.e., they are in the same equivalence class). [TODO: show that it achieves identifiability condition]

### 8.7.2 Computing $z_n$ and the prior variance

Now that we have defined our prior and kernel, we can compute the “mini-normalization constants”,  $z_1, \dots, z_n$ , given by (8.65). These quantities represent the expected covariance of the likelihood of a latent structure  $\theta'$ , with the likelihood of another structure drawn from the prior.

If matrices  $\mathbf{Z}$  and  $\mathbf{Z}'$  have  $K$  and  $K'$  columns respectively, then combining (8.68) and (8.70), we have:

$$z_n(\mathbf{Z}) = \sum_{\mathbf{Z}'} k(\mathbf{Z}, \mathbf{Z}') p(\mathbf{Z}' | \alpha) = \sum_{\mathbf{Z}'} \left[ \sum_{n=1}^N \sum_{k=1}^K \sum_{k'=1}^{K'} \mathbf{Z}_{n,k} \mathbf{Z}'_{n,k'} \right] \left[ \prod_{k^*=1}^K P(\mathbf{Z}'_{:,k^*} | \alpha) \right] \quad (8.71)$$

$$= \sum_{k=1}^K \sum_{n=1}^N \mathbf{Z}_{n,k} \frac{\alpha}{1 + \frac{\alpha}{K'}} = \alpha \sum_{k=1}^K \sum_{n=1}^N \mathbf{Z}_{n,k} \quad \text{as } K' \rightarrow \infty \quad (8.72)$$

See the supplementary material for longer derivations. We can interpret (8.72) as saying that the prior covariance of the likelihood  $p(\mathbf{X}|\mathbf{Z})$  with a randomly drawn likelihood  $p(\mathbf{X}|\mathbf{Z}')$  is proportional to the number of ones in  $\mathbf{Z}$ .

Next we compute the prior variance (8.66), which follows a similar derivation. This is the expected variance of  $Z$  before we have seen any data.

$$V_{\text{prior}} = \sum_{\mathbf{Z}'} \sum_{\mathbf{Z}} p(\mathbf{Z} | \alpha) k(\mathbf{Z}, \mathbf{Z}') p(\mathbf{Z}' | \alpha) = \frac{\alpha^2 N}{1 + \frac{\alpha}{K}} = \alpha^2 N \quad \text{as } K \rightarrow \infty \quad (8.73)$$

This form is intuitive: it scales with the expected number of ones per row,  $\alpha$ . As  $\alpha$  or  $N$  approach zero, the likelihood surface can be expected to become less varied, since there will be fewer ways that individual samples of  $\mathbf{Z}$  can differ.

## 8.8 Infinite Mixture Models

In this section we develop a more general example, placing a kernel between mixture distributions. This will allow us to perform model-based inference in Dirichlet process mixture models.

A finite mixture model with  $k$  components gives a distribution over the observations  $x$  as follows:  $p(x|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^k \pi_i p(x|\theta_i)$  where  $\theta_i$  represents a single set of mixture parameters. By specifying the form of  $\theta$ , we can perform inference in a wide variety of infinite mixture models, such as an latent Dirichlet allocation-type model, infinite mixture of regressors, or a mixture of Gaussians.

We will divide our derivation into two parts. First, we will define a kernel between multinomial distributions, and derive (8.65) and (8.66) for this kernel, leaving unspecified the kernel between individual mixture elements. Then, we will continue the derivation for an infinite mixture of Gaussians.

### 8.8.1 A Kernel Between Multinomial Distributions

Here we define a kernel between multinomial distributions, possibly of different dimension. Here,  $\boldsymbol{\pi}$  represents weights which sum to one, and  $\theta$  the atoms of the multinomial distribution.

$$k(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\pi}', \boldsymbol{\theta}') = \sum_i^{n_\theta} \sum_j^{n_{\theta'}} \pi_i \pi'_j k_a(\theta_i, \theta'_j) \quad (8.74)$$

where  $k_a(\theta, \theta')$  specifies the covariance between individual atoms of the distribution. Changing the order of mixture components, or splitting a given mixture component among two identical atoms, will not affect the value of this covariance function. If  $k_a(\theta_i, \theta'_j)$  is a Mercer kernel, then so is (8.74).

If our mixture  $\theta$  has  $n_\theta$  components, then

$$\begin{aligned} z(\boldsymbol{\pi}, \boldsymbol{\theta}) &= \iint k(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\pi}', \boldsymbol{\theta}') p(\boldsymbol{\theta}') p(\boldsymbol{\pi}') d\boldsymbol{\theta}' d\boldsymbol{\pi}' = \iint \left[ \sum_{i=1}^{n_\theta} \sum_{j=1}^{n_{\theta'}} \pi_i \pi'_j k(\theta_i, \theta'_j) \right] \left[ \prod_a^{n_{\theta'}} p(\theta'_a) \right] p(\boldsymbol{\pi}') d\boldsymbol{\theta}' d\boldsymbol{\pi}' \\ &= \int \sum_{i=1}^{n_\theta} \sum_{j=1}^{n_{\theta'}} \pi_i \pi'_j \underbrace{\int k(\theta_i, \theta'_j) p(\theta'_j) d\theta'_j}_{z_a(\theta_i)} p(\boldsymbol{\pi}') d\boldsymbol{\pi}' = \sum_{i=1}^{n_\theta} \pi_i z_a(\theta_i) \end{aligned} \quad (8.75)$$

$$\mathbb{V}_{\text{prior}} [Z_{f,p}] = \iint z(\boldsymbol{\pi}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\pi}) d\boldsymbol{\theta} d\boldsymbol{\pi} = \underbrace{\int z(\theta_i) p(\theta_i) d\theta_i}_{V_a} \underbrace{\int p(\boldsymbol{\pi}) \sum_{i=1}^{n_\theta} \pi_i d\boldsymbol{\pi}}_{\text{sums to one}} = V_a \quad (8.76)$$

where  $V_a = \iint p(\boldsymbol{\theta}) k_a(\boldsymbol{\theta}, \boldsymbol{\theta}') p(\boldsymbol{\theta}') d\boldsymbol{\theta}' d\boldsymbol{\theta}$  is the prior variance of  $k_a$ , which will be defined below.

Perhaps surprisingly, neither  $z_n$  nor  $V_{\text{prior}}$  depend on the number of proposed mixture components. This means that we are free to take the infinite limit  $n_\theta \rightarrow \infty$ . Perhaps this makes sense: The likelihood does not change if we divide up our clusters to give equivalent mixtures but having more components. These quantities also do not depend on the form of the prior over mixture components  $p(\boldsymbol{\pi})$ , nor the prior over individual components  $p(\boldsymbol{\theta})$ , as long as it factorizes over components.

### 8.8.2 Infinite Mixture of Gaussians

The above kernel between mixtures can be used for inference in a wide variety of infinite mixture models. For simplicity, in this paper we will use as an example the infinite mixture of Gaussians Rasmussen (2000):

$$p(x|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^k \pi_i p(x|\theta_i) = \sum_{i=1}^k \pi_i \mathcal{N}(y|\mu_i, \Sigma_i) \quad (8.77)$$

where  $\theta_i = \{\mu_i, \Sigma_i\}$  represent the parameters of a single Gaussian. To complete our example, all that remains is to specify a kernel  $k_a$  between individual mixture components, and to compute  $z_k$  and  $V_a$ . For simplicity, we will specify a Gaussian kernel between densities, and assume that the variance of each mixture component is identical:

$$k(\mu, \Sigma, \mu', \Sigma') = \mathcal{N}(\mu|\mu', \Sigma_k) \quad (8.78)$$

where the entries of  $\Sigma_k$  are kernel parameters. Next, we give (8.65) and (8.66) for this kernel:

$$z_k(\mu_i, \Sigma_i) = \iint k(\mu_i, \Sigma_i, \mu'_j, \Sigma'_j) p(\mu'_j, \Sigma'_j) d\mu'_j d\Sigma'_j = \mathcal{N}(\mu_i|\lambda, \Sigma_k + \Sigma_p) \quad (8.79)$$

$$V_a = \int z_k(\mu_i, \Sigma_i) p(\mu_i, \Sigma_i) d\mu_i d\Sigma_i = \mathcal{N}(0|0, \Sigma_k + 2\Sigma_p) \quad (8.80)$$

Note that the prior variance  $V_k$  depends only on the shape of the prior  $\Sigma_p$ , and not its location.

## 8.9 Related Work

String kernels ?. Graph kernels. Tree-structured kernels. Sequence kernels. A review of kernels in structured domains can be found in ?.

## 8.10 Experiments

**Integrating Kernel Hyperparameters** One complicating issue of inference using BQ is how to set kernel hyperparameters. In ?, these were set by maximum likelihood. In our experiments, hyperparameters were integrated out numerically, except for the *output variance* ( a scale factor in front of the kernel ) which is possible to integrate out

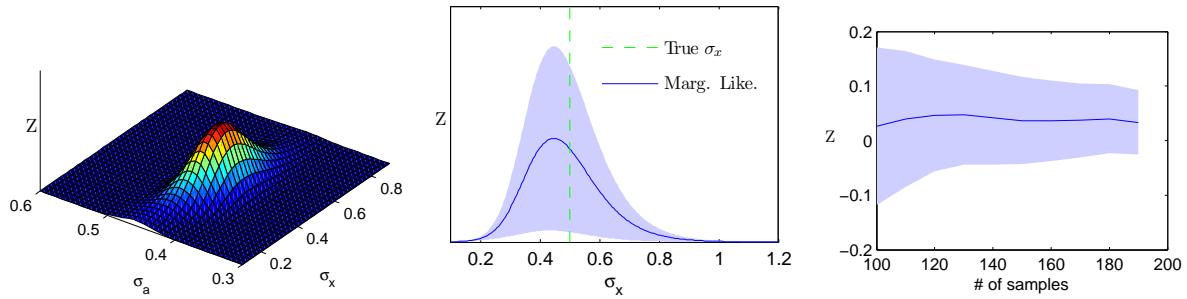


Figure 8.4 Computing marginal likelihoods with BQ. Left: The marginal likelihood as a function of two nuisance parameters. Center: A slice of the 2-D marginal likelihood function, at  $\sigma_A = 0.4$ . The true value of  $\sigma_X$  lies roughly in the center of the likelihood function. Right: The marginal likelihood as a function of the number of evaluations of the likelihood function. The shaded error represents uncertainty about the value of the marginal likelihood.

in closed form, giving a final posterior variance of  $\sigma_2 = \frac{1}{N} [\mathbf{y}^t \mathbf{K}^{-1} \mathbf{y}] [V_k - \mathbf{z}_k \mathbf{K}^{-1} \mathbf{z}]$ . For a derivation, see the supplementary material.

### 8.10.1 IBP Experiments

We used collapsed IBP sampling code from ? to obtain samples and likelihood values. We used data from ?, where the true value of  $\sigma_x = 0.5$ . We set the number of datapoints to be small ( $N = 25$ ), since this is a regime where VB is known to perform poorly ?.

Figure 8.4 demonstrates the use of BQ, showing that a set of samples gathered under one parameter setting can be used to make inferences the likelihood of other parameter settings. This allows us to use incorrect samplers, use samples from the burn-in period, etc. For example, the likelihood function (8.69) has two “nuisance parameters”,  $\sigma_X$  and  $\sigma_A$ . In [cite Finale], these parameters are simply set in an ad-hoc way. This may be acceptable, but using MCMC, it is not clear how to tell whether the answer is sensitive to these nuisance parameters without re-running the chain under several different settings. Figure 8.4 shows that BQ allows one to run a post-hoc sensitivity analysis to check whether extra parameters are important to the analysis. In addition, we recover an estimate of the certainty of our analysis, indicating whether or not the existing samples are sufficient to draw strong conclusions.

### 8.10.2 DP Mixture Experiments

Figure 8.4 demonstrates the use of BQ to examine sensitivity to prior parameters.

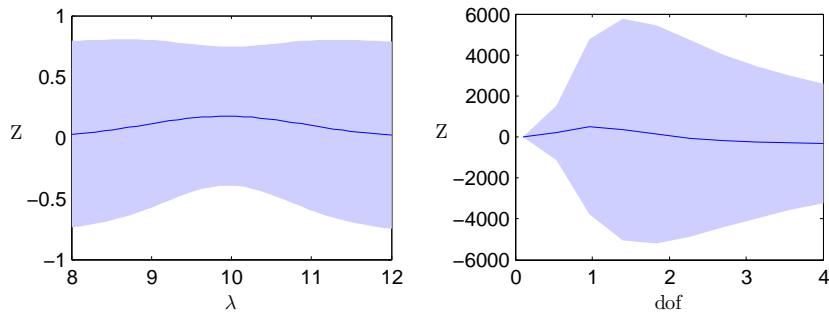


Figure 8.5 A demonstration of computing marginal likelihoods. Left: The marginal likelihood versus the prior mean. Right: Marginal likelihood versus the degrees of freedom of the mixture components.  $dof = 1$  is Cauchy,  $dof = 2$  is Student's t,  $dof = \infty$  is Gaussian. The shaded error represents uncertainty about the value of the marginal likelihood.

Figures 8.4 and 8.5 show not only the marginal likelihood of different parameter settings, but also the marginal variance of the likelihood functions. We must distinguish between two types of uncertainty represented here. First, the shape of the likelihood function indicates our posterior uncertainty over its parameter, given the data. Second, the GP posterior uncertainty in the likelihood function (represented by the shaded areas) represents our uncertainty about this likelihood function. Our uncertainty about the likelihood function can be made arbitrarily small by continuing to sample the likelihood function. However, we may still remain uncertain about the value of the latent parameter.

Code to produce all experiments will be made available at the authors' website.

## 8.11 Discussion

Nuisance parameters such as  $\sigma_X$  could also be dealt with in the BQ by adding them to the kernel and the domain of integration if desired.

### 8.11.1 Appropriateness of the GP prior

Placing a GP prior on the likelihood function allows us to take advantage of our knowledge of the smoothness of this function. However, there is ample reason to believe that the GP prior is inappropriate for modeling likelihood functions. In ? it is suggested to place the GP on the log-likelihood function, which would presumably make the additive form of the kernels given in the paper much more appropriate. This was done

by [cite BQR paper], and resulted in better performance, at the cost of a much more complicated inference procedure.

As is generally true for Bayesian methods, there exist many cases where BQ will underestimate its uncertainty. Our response to this objection is that any uncertainty estimate is better than none at all. In our experiments, we observed cases in which the model's posterior uncertainty is significant, alerting us to the fact that we have not yet observed enough about the likelihood function to be certain about its shape. In the case of MCMC, this sort of uncertainty estimate is typically unavailable. That is to say, our model cannot account for ‘unknown unknowns’, but it can at least account for ‘known unknowns’, which is a step up from the point estimates provided by MCMC.

## 8.12 Conclusions

We have extended Bayesian quadrature to infinite, structured domains, and demonstrated that this method can be used for inference in nonparametric models. We have given necessary conditions for convergence and examples of how to construct kernels which take advantage of the many symmetries of typical likelihood functions. We demonstrated some properties of this method, which include uncertainty estimates, flexibility in sampling methods, and the ability to re-use samples from on setting to perform post-hoc sensitivity analysis of nuisance parameters.

## 8.13 Gaussian Process Posteriors

Imagine we have a GP posterior distribution over functions, after having conditioned on data points  $\mathbf{X}, \mathbf{f}(\mathbf{X})$ .

Our posterior distribution is given by:

$$p(\mathbf{f}(\mathbf{x}_*) | \mathbf{X}, \mathbf{f}(\mathbf{X})) = \mathcal{N}\left(\mathbf{f}(\mathbf{x}_*) \mid \bar{\mathbf{f}}(\mathbf{x}_*), \text{cov}(\mathbf{x}_*, \mathbf{x}'_*)\right) \quad (8.81)$$

where

$$\bar{\mathbf{f}}(\mathbf{x}_*) = \mathbb{E}_{\text{GP}} [f(\mathbf{x}_*)] = k(\mathbf{x}_*, \mathbf{X}) K^{-1} \mathbf{f}(\mathbf{X}) \quad (8.82)$$

$$\text{cov}(\mathbf{x}_*, \mathbf{x}'_*) = \mathbb{V}_{\text{GP}} [f(\mathbf{x}_*)] = k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X}) K^{-1} k(\mathbf{X}, \mathbf{x}_*) \quad (8.83)$$

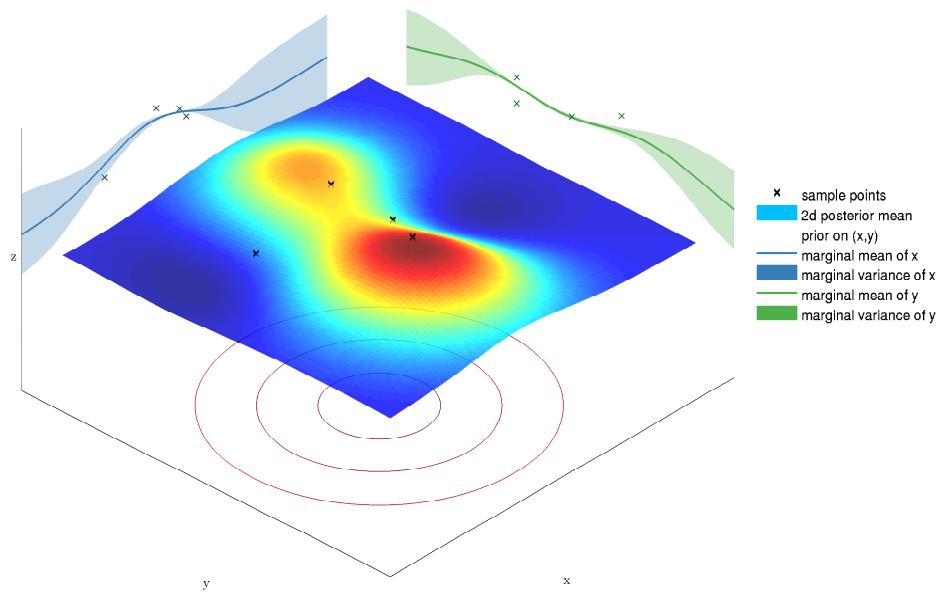


Figure 8.6 A 2D GP posterior, and its 1D marginals.

Often, we are interested in the posterior distribution over this function, integrating over several input dimensions. Let us redefine our function to be over a set of variables  $\mathbf{x}$ , which we are interested in, and nuisance variables  $\mathbf{y}$ , which we wish to integrate out.

$$p(\mathbf{f}(\mathbf{x}_*, \mathbf{y}_*) | \mathbf{X}, \mathbf{Y}, \mathbf{f}(\mathbf{X}, \mathbf{Y})) = \mathcal{N}\left(\mathbf{f}(\mathbf{x}_*, \mathbf{y}_*) \mid \bar{\mathbf{f}}(\mathbf{x}_*, \mathbf{y}_*), \text{cov}(\mathbf{x}_*, \mathbf{y}_*, \mathbf{x}'_*, \mathbf{y}'_*)\right) \quad (8.84)$$

where

$$\bar{\mathbf{f}}(\mathbf{x}_*, \mathbf{y}_*) = \mathbb{E}_{\text{GP}} [f(\mathbf{x}_*, \mathbf{y}_*)] = k(\mathbf{x}_*, \mathbf{y}_*, \mathbf{X}, \mathbf{Y}) K^{-1} \mathbf{f}(\mathbf{X}, \mathbf{Y}) \quad (8.85)$$

$$\text{cov}(\mathbf{x}_*, \mathbf{x}_*) = \mathbb{V}_{\text{GP}} [f(\mathbf{x}_*, \mathbf{y}_*)] = k(\mathbf{x}_*, \mathbf{y}_*, \mathbf{x}_*, \mathbf{y}_*) - k(\mathbf{x}_*, \mathbf{y}_*, \mathbf{X}, \mathbf{Y}) K^{-1} k(\mathbf{X}, \mathbf{Y}, \mathbf{x}_*, \mathbf{y}_*) \quad (8.86)$$

We are interested in the marginal

$$p(\mathbf{f}(\mathbf{x}_*)|\mathbf{X}, \mathbf{Y}, \mathbf{f}(\mathbf{X}, \mathbf{Y})) = \int p(\mathbf{f}(\mathbf{x}_*, \mathbf{y})|\mathbf{X}, \mathbf{Y}, \mathbf{f}(\mathbf{X}, \mathbf{Y}))d\mathbf{y} \quad (8.87)$$

$$= \int \mathcal{N}\left(\mathbf{f}(\mathbf{x}_*, \mathbf{y}_*) \mid \bar{\mathbf{f}}(\mathbf{x}_*, \mathbf{y}_*), \text{cov}(\mathbf{x}_*, \mathbf{y}_*, \mathbf{x}'_*, \mathbf{y}'_*)\right) d\mathbf{y} \quad (8.88)$$

### 8.13.1 Marginal Mean Function

Here we work out the posterior marginal mean function  $\mathbb{E}[\mathbf{f}(\mathbf{x}_*)]$ :

$$\mathbb{E}_{p(f(\mathbf{x}_*)|\mathbf{X}, \mathbf{Y}, \mathbf{f}(\mathbf{X}, \mathbf{Y}))} [\mathbf{f}(\mathbf{x}_*)] = \iint f(\mathbf{x}_*, \mathbf{y}) p(f(\mathbf{x}_*, \mathbf{y})|\mathbf{X}, \mathbf{Y}, \mathbf{f}(\mathbf{X}, \mathbf{Y})) p(\mathbf{y}|\mathbf{x}_*) d\mathbf{y} df \quad (8.89)$$

$$= \int \mathbb{E}_{p(f(\mathbf{x}_*, \mathbf{y})|\mathbf{X}, \mathbf{Y}, \mathbf{f}(\mathbf{X}, \mathbf{Y}))} [f(\mathbf{x}_*, \mathbf{y})] p(\mathbf{y}|\mathbf{x}_*) d\mathbf{y} \quad (8.90)$$

$$= \int \bar{\mathbf{f}}(\mathbf{x}_*, \mathbf{y}) p(\mathbf{y}) d\mathbf{y} \quad (8.91)$$

$$= \int k(\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \mathbf{Y}) \underbrace{K^{-1} \mathbf{f}(\mathbf{X}, \mathbf{Y})}_{\beta} p(\mathbf{y}|\mathbf{x}_*) d\mathbf{y} \quad (8.92)$$

$$= \sum_i \beta_i \int k(\mathbf{x}_*, \mathbf{y}, \mathbf{X}_i, \mathbf{Y}_i) p(\mathbf{y}|\mathbf{x}_*) d\mathbf{y} \quad (8.93)$$

If we choose the kernel function and prior such that

$$k(\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}') = h\mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{x}' \\ \mathbf{y}' \end{bmatrix}, \begin{bmatrix} I\ell_{\mathbf{x}} & 0 \\ 0 & I\ell_{\mathbf{y}} \end{bmatrix}\right) \quad (8.94)$$

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{bmatrix}\right) \quad (8.95)$$

then the conditional prior on  $p(\mathbf{y}|\mathbf{x}_*)$  is:

$$p(\mathbf{y}|\mathbf{x}_*) = \mathcal{N}\left(\mathbf{y} \mid \mu_{\mathbf{y}|\mathbf{x}_*}, \Sigma_{\mathbf{y}|\mathbf{x}_*}\right) \quad (8.96)$$

$$\mu_{\mathbf{y}|\mathbf{x}_*} = \mu_{\mathbf{y}} + \Sigma_{\mathbf{xy}} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{x}_* - \mu_{\mathbf{x}}) \quad (8.97)$$

$$\Sigma_{\mathbf{y}|\mathbf{x}_*} = \mu_{\mathbf{y}} + \Sigma_{\mathbf{yx}} \Sigma_{\mathbf{xx}}^{-1} \Sigma_{\mathbf{xy}} \quad (8.98)$$

then we can obtain a closed-form solution for Equation (8.93):

$$\int k(\mathbf{x}_*, \mathbf{y}, \mathbf{X}_i, \mathbf{Y}_i) p(\mathbf{y}) d\mathbf{y} = \int h \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{x}' \\ \mathbf{y}' \end{bmatrix}, \begin{bmatrix} I\ell_{\mathbf{x}} & 0 \\ 0 & I\ell_{\mathbf{y}} \end{bmatrix}\right) \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} \quad (8.99)$$

$$= \int h \mathcal{N}(\mathbf{x} | \mathbf{X}_i, I\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{y} | \mathbf{Y}_i, I\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} \quad (8.100)$$

$$= h \mathcal{N}(\mathbf{x} | \mathbf{X}_i, I\ell_{\mathbf{x}}) \int \mathcal{N}(\mathbf{y} | \mathbf{Y}_i, I\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} \quad (8.101)$$

$$= h \mathcal{N}(\mathbf{x} | \mathbf{X}_i, I\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y}_i | \mu_{\mathbf{y}}, I\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) \quad (8.102)$$

Thus our final form for  $\mathbb{E} [\mathbf{f}(\mathbf{x}_*)]$  is:

$$\mathbb{E}_{p(\mathbf{f}(\mathbf{x}_*) | \mathbf{X}, \mathbf{Y}, \mathbf{f}(\mathbf{X}, \mathbf{Y}))} [\mathbf{f}(\mathbf{x}_*)] = \sum_i \beta_i \int k(\mathbf{x}_*, \mathbf{y}, \mathbf{X}_i, \mathbf{Y}_i) p(\mathbf{y}) d\mathbf{y} \quad (8.103)$$

$$= \sum_i \beta_i h \mathcal{N}(\mathbf{x}_* | \mathbf{X}_i, I\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y}_i | \mu_{\mathbf{y}}, I\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) \quad (8.104)$$

### 8.13.2 Marginal Variance Function

One major advantage of model-based approaches to integration is that they provide a value for the uncertainty in the given estimate. Here, we derive the marginal covariance function  $\mathbb{V} [f(\mathbf{x}_*), f(\mathbf{x}'_*)]$ :

$$\text{Cov} [f(\mathbf{x}_\star), f(\mathbf{x}'_\star)] = \mathbb{E}_{p(f)} \left[ (f(\mathbf{x}_\star) - \bar{f}(\mathbf{x}_\star)) (f(\mathbf{x}'_\star) - \bar{f}(\mathbf{x}'_\star)) \right] \quad (8.105)$$

$$= \mathbb{E}_{p(f)} \left[ \left( \int f(\mathbf{x}_\star, \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \int \bar{f}(\mathbf{x}_\star, \mathbf{y}') p(\mathbf{y}') d\mathbf{y}' \right) \right] \quad (8.106)$$

$$\times \left( \int f(\mathbf{x}'_\star, \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \int \bar{f}(\mathbf{x}'_\star, \mathbf{y}') p(\mathbf{y}') d\mathbf{y}' \right) \quad (8.107)$$

$$= \int \left( \int f(\mathbf{x}_\star, \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \int \bar{f}(\mathbf{x}_\star, \mathbf{y}') p(\mathbf{y}') d\mathbf{y}' \right) \quad (8.108)$$

$$\times \left( \int f(\mathbf{x}'_\star, \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \int \bar{f}(\mathbf{x}'_\star, \mathbf{y}') p(\mathbf{y}') d\mathbf{y}' \right) p(f) \quad (8.109)$$

$$= \iiint [f((\mathbf{x}_\star, \mathbf{y}) - \bar{f}(\mathbf{x}_\star, \mathbf{y}')] [f((\mathbf{x}'_\star, \mathbf{y}) - \bar{f}(\mathbf{x}'_\star, \mathbf{y}')] p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' p(f) df \quad (8.110)$$

$$= \iiint [f((\mathbf{x}_\star, \mathbf{y}) - \bar{f}(\mathbf{x}_\star, \mathbf{y}')] [f((\mathbf{x}'_\star, \mathbf{y}') - \bar{f}(\mathbf{x}'_\star, \mathbf{y}')] p(f) df p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' \quad (8.111)$$

$$= \iint \text{Cov} [f((\mathbf{x}_\star, \mathbf{y}), f((\mathbf{x}'_\star, \mathbf{y}')) p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' \quad (8.112)$$

$$= \iint [k(\mathbf{x}_\star, \mathbf{y}, \mathbf{x}_\star, \mathbf{y}') - k(\mathbf{x}_\star, \mathbf{y}, \mathbf{X}, \mathbf{Y}) K^{-1} k(\mathbf{X}, \mathbf{Y}, \mathbf{x}_\star, \mathbf{y}')] p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' \quad (8.113)$$

$$= \iint k(\mathbf{x}_\star, \mathbf{y}, \mathbf{x}_\star, \mathbf{y}') p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' - \iint k(\mathbf{x}_\star, \mathbf{y}, \mathbf{X}, \mathbf{Y}) K^{-1} k(\mathbf{X}, \mathbf{Y}, \mathbf{x}_\star, \mathbf{y}') p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' \quad (8.114)$$

$$= t_1(\mathbf{x}_\star, \mathbf{x}'_\star) - t_2(\mathbf{x}_\star, \mathbf{x}'_\star) \quad (8.115)$$

Again, if we choose the kernel function and prior that consist of scaled Normal distributions, then we can obtain a closed-form solution for Equation (8.115). Here we consider

the first term,  $t_1$ :

$$t_1(\mathbf{x}_*, \mathbf{x}'_*) = \iint k(\mathbf{x}_*, \mathbf{y}, \mathbf{x}_*, \mathbf{y}') p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' \quad (8.116)$$

$$= \iint h \mathcal{N} \left( \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{x}' \\ \mathbf{y}' \end{bmatrix}, \begin{bmatrix} I\ell_{\mathbf{x}} & 0 \\ 0 & I\ell_{\mathbf{y}} \end{bmatrix} \right) \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{y}'|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} d\mathbf{y}' \quad (8.117)$$

$$= \iint h \mathcal{N}(\mathbf{x}_*|\mathbf{x}'_*, I\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{y}|\mathbf{y}', I\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{y}'|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} d\mathbf{y}' \quad (8.118)$$

$$= h \mathcal{N}(\mathbf{x}_*|\mathbf{x}'_*, I\ell_{\mathbf{x}}) \iint \mathcal{N}(\mathbf{y}|\mathbf{y}', I\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{y}'|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} d\mathbf{y}' \quad (8.119)$$

$$= h \mathcal{N}(\mathbf{x}_*|\mathbf{x}'_*, I\ell_{\mathbf{x}}) \int \mathcal{N}(\mathbf{y}'|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \int \mathcal{N}(\mathbf{y}|\mathbf{y}', I\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} d\mathbf{y}' \quad (8.120)$$

$$= h \mathcal{N}(\mathbf{x}_*|\mathbf{x}'_*, I\ell_{\mathbf{x}}) \int \mathcal{N}(\mathbf{y}'|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{y}'|\mu_{\mathbf{y}}, I\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) d\mathbf{y}' \quad (8.121)$$

$$= h \mathcal{N}(\mathbf{x}_*|\mathbf{x}'_*, I\ell_{\mathbf{x}}) \mathcal{N}(\mu_{\mathbf{y}}|\mu_{\mathbf{y}}, I\ell_{\mathbf{y}} + 2\Sigma_{\mathbf{y}}) \quad (8.122)$$

$$= h \mathcal{N}(\mathbf{x}_*|\mathbf{x}'_*, I\ell_{\mathbf{x}}) \mathcal{N}(0|0, I\ell_{\mathbf{y}} + 2\Sigma_{\mathbf{y}}) \quad (8.123)$$

$t_1(\mathbf{x}_*, \mathbf{x}'_*)$  is called the *prior covariance*, and does not depend on the observed function values.

Next, we consider the second term,  $t_2$ :

$$t_2(\mathbf{x}_*, \mathbf{x}'_*) = \quad (8.124)$$

$$= \iint k(\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \mathbf{Y}) K^{-1} k(\mathbf{X}, \mathbf{Y}, \mathbf{x}'_*, \mathbf{y}') p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' \quad (8.125)$$

$$= \sum_i \sum_j K_{ij}^{-1} \iint k(\mathbf{x}_*, \mathbf{y}, \mathbf{X}_i, \mathbf{Y}_i) k(\mathbf{X}_j, \mathbf{Y}_j, \mathbf{x}'_*, \mathbf{y}') p(\mathbf{y}) p(\mathbf{y}') d\mathbf{y} d\mathbf{y}' \quad (8.126)$$

$$= \sum_i \sum_j K_{ij}^{-1} \iint h\mathcal{N}(\mathbf{x}_* | \mathbf{X}_i, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{y} | \mathbf{Y}_i, \text{I}\ell_{\mathbf{y}}) h\mathcal{N}(\mathbf{x}'_* | \mathbf{X}_j, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{y}' | \mathbf{Y}_j, \text{I}\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{y}' | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} d\mathbf{y}' \quad (8.127)$$

$$= \sum_i \sum_j K_{ij}^{-1} h\mathcal{N}(\mathbf{x}_* | \mathbf{X}_i, \text{I}\ell_{\mathbf{x}}) h\mathcal{N}(\mathbf{x}'_* | \mathbf{X}_j, \text{I}\ell_{\mathbf{x}}) \iint \mathcal{N}(\mathbf{y} | \mathbf{Y}_i, \text{I}\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y}' | \mathbf{Y}_j, \text{I}\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{y}' | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} d\mathbf{y}' \quad (8.128)$$

$$= \sum_i \sum_j K_{ij}^{-1} h\mathcal{N}(\mathbf{x}_* | \mathbf{X}_i, \text{I}\ell_{\mathbf{x}}) h\mathcal{N}(\mathbf{x}'_* | \mathbf{X}_j, \text{I}\ell_{\mathbf{x}}) \int \mathcal{N}(\mathbf{y} | \mathbf{Y}_i, \text{I}\ell_{\mathbf{y}}) \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) d\mathbf{y} \int \mathcal{N}(\mathbf{y}' | \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{y}' | \mathbf{Y}_j, \text{I}\ell_{\mathbf{y}}) d\mathbf{y}' \quad (8.129)$$

$$= \sum_i \sum_j K_{ij}^{-1} h\mathcal{N}(\mathbf{x}_* | \mathbf{X}_i, \text{I}\ell_{\mathbf{x}}) h\mathcal{N}(\mathbf{x}'_* | \mathbf{X}_j, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y}_i | \mu_{\mathbf{y}}, \text{I}\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) \mathcal{N}(\mathbf{Y}_j | \mu_{\mathbf{y}}, \text{I}\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) \quad (8.130)$$

$$= \sum_i \sum_j K_{ij}^{-1} h\mathcal{N}(\mathbf{x}_* | \mathbf{X}_i, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y}_i | \mu_{\mathbf{y}}, \text{I}\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) h\mathcal{N}(\mathbf{x}'_* | \mathbf{X}_j, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y}_j | \mu_{\mathbf{y}}, \text{I}\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) \quad (8.131)$$

$$= h\mathcal{N}(\mathbf{x}_* | \mathbf{X}, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y} | \mu_{\mathbf{y}}, \text{I}\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) K^{-1} h\mathcal{N}(\mathbf{x}'_* | \mathbf{X}, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y} | \mu_{\mathbf{y}}, \text{I}\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) \quad (8.132)$$

$$= \mathbf{z}(\mathbf{x}_*) K^{-1} \mathbf{z}(\mathbf{x}'_*) \quad (8.133)$$

where

$$z_i(\mathbf{x}) = h\mathcal{N}(\mathbf{x} | \mathbf{X}_i, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(\mathbf{Y}_i | \mu_{\mathbf{y}}, \text{I}\ell_{\mathbf{y}} + \Sigma_{\mathbf{y}}) \quad (8.134)$$

Thus our final form for  $\text{Cov}[f(\mathbf{x}_*), f(\mathbf{x}'_*)]$  is:

$$\text{Cov}_{p(f(\mathbf{x}_*) | \mathbf{X}, \mathbf{Y}, f(\mathbf{x}, \mathbf{y}))} [f(\mathbf{x}_*), f(\mathbf{x}'_*)] = t_1(\mathbf{x}_*, \mathbf{x}'_*) - t_2(\mathbf{x}_*, \mathbf{x}'_*) \quad (8.135)$$

$$= h\mathcal{N}(\mathbf{x}_* | \mathbf{x}'_*, \text{I}\ell_{\mathbf{x}}) \mathcal{N}(0 | 0, \text{I}\ell_{\mathbf{y}} + 2\Sigma_{\mathbf{y}}) - \mathbf{z}(\mathbf{x}_*) K^{-1} \mathbf{z}(\mathbf{x}'_*) \quad (8.136)$$

## 8.14 Objective Function

When choosing points at which to evaluate  $f(\mathbf{x})$  in order to learn more about the integrand, there is a question about which objective function to minimize. Standard sampling for BQ attempts to minimize the expected variance in our integral,  $Z$ . However, if we are only interested in the marginal distribution over some variable, then we might be afraid that minimizing the variance in  $Z$  is only tangentially related to what we care about.

Here, we show that minimizing the variance in  $Z$  is equivalent to minimizing the variance of the marginal function, weighted by the prior  $p(\mathbf{x})$ .

...

Note that since our posterior over  $Z$  is a Gaussian, minimizing the variance in  $Z$  is equivalent to minimizing the entropy of our distribution over  $Z$ .

# Chapter 9

## Discussion

This chapter articulates some of the questions raised by the work in this thesis, and attempts to relate kernel-based model-building to the larger project of automating statistics and machine learning. We conclude with a summary of the contributions of this thesis.

### 9.1 Structured versus unstructured models

One question left unanswered by this thesis is when to prefer the structured, kernel-based models described chapters 2, 3 and 6 to the relatively unstructured deep GP models described in chapter 5. In this section, we consider some advantages and disadvantages of the two approaches.

- **Difficulty of optimization.** The discrete nature of the search over composite kernel structures can be seen as both a blessing and a curse. Certainly, a mixed discrete and continuous search space requires more complex optimization procedures compared to the continuous-only optimization possible in deep GPs.

However, the discrete nature of the space of composite kernels offers the possibility of learning heuristics to suggest which types of structure to add. For example, finding periodic structure or growing variance in the residuals of a model suggests adding periodic or linear components to the kernel, respectively. It is not clear whether such heuristics can easily be found for optimizing the variational parameters of a deep GP, although

- **Long-range extrapolation.** Another question is whether the inductive bias of deep GPs can enable the sorts of long-range extrapolation shown in chapters 2

and 3. As an example, consider the problem of extrapolating a periodic function. A deep GP could learn a latent representation similar to that of the periodic kernel, projecting into a basis equivalent to  $[\sin(x), \cos(x)]$  in the first hidden layer. However, to extrapolate a periodic function, the sin and cos functions would have to continue to repeat beyond the range of the training data, which would not happen if each layer assumed only local smoothness.

One obvious possibility is to marry the two approaches, learning deep GPs with structured kernels. However, we may lose some of the advantages of interpretability by this approach.

Another point to consider is that, in high dimensions, the line between interpolation and extrapolation is blurred. Learning a suitable representation of the data manifold may be sufficient for most purposes.

- **Ease of interpretation.** Historically, the statistics community has put more emphasis on the interpretability and meaning of models than the machine learning community, which has focused more on predictive performance. To automate the practice of statistics, developing model-description procedures for powerful model classes seems to be a necessary step.

In general, such descriptions would seem to require some sort of decomposition of a model into discrete components. For example, chapter 4 showed that composite kernels allow automatic visualization and description of low-dimensional structure.

On the other hand, [Damianou and Lawrence \(2013\)](#) showed that deep GP-LVMs allow summarization of high-dimensional structure through sampling from the posterior, examining the dimension of each latent layer, visualizing latent coordinates, and examining how the predictive distribution changes as one moves in the latent space. Perhaps more sophisticated procedures would also allow intelligible descriptions of such models.

The warped mixture model of chapter 7 represents a compromise between these two approaches, combining a discrete clustering model with an unstructured warping function. However, the explicit clustering model may be unnecessary: the results of [Damianou and Lawrence \(2013\)](#) suggest that clustering can be automatically accomplished by a sufficiently deep, unstructured GP.

## 9.2 Approaches to automating model construction

This thesis is part of a larger push to automate the practice of model building and inference. Broadly speaking, this problem is being attacked from two directions.

From the top-down, the probabilistic programming community is developing automatic inference engines for extremely broad classes of models (Goodman et al., 2008; Liang et al., 2010; Mansinghka et al., 2014) such as the class of all computable distributions (Li and Vitányi, 1997; Solomonoff, 1964). As discussed in section 3.1, model construction can be seen as search through an open-ended model class. Universal search strategies have been constructed for the space of all computable distributions (Hutter, 2002; Levin, 1973; Schmidhuber, 2002), but they remain impractically slow.

The alternative, bottom-up approach is to design procedures which extend and combine existing model classes, for which relatively efficient inference algorithms are known. For example, Grosse (2014) built an open-ended language of matrix decomposition models and a corresponding compositional language of relatively efficient approximate inference algorithms. This approach makes inference feasible for models in the language, but extending the language requires developing new inference algorithms. As another example, Steinruecken (2014) showed how to compose inference algorithms for arbitrary sequence models. The language of models proposed in chapter 3 is an example of this bottom-up approach, and has the same benefits and limitations.

If the building-blocks being composed are sufficiently powerful, the line between these two approaches becomes blurred. For example, we deep generative models (Adams et al., 2010; Bengio et al., 2013; Damianou and Lawrence, 2013; Rippel and Adams, 2013; Salakhutdinov and Hinton, 2009) could be considered an example of the bottom-up approach, since they compose both model and inference into individual layers. However, large neural nets can capture enough different types of structure that they could be seen as an example of the universalist top-down approach.

## 9.3 Summary of contributions

The main contribution of this thesis was to show how to automate the construction of interpretable nonparametric regression models using Gaussian processes. This was done in several parts: First, chapter 2 presented a systematic overview of kernel construction techniques, as well as properties of the resulting GP priors. Next, chapter 3 showed the viability of a bread-first search over an open-ended space of kernels, and showed that

the corresponding GP models could be automatically decomposed into diverse parts illustrating the structure found in the data. Chapter 4 showed that each part of a kernel can be described modularly, allowing automatically written text to be included in detailed reports describing GP models. An example report is included in appendix D. Together, these chapters describe the beginnings of an “automatic statistician”, capable of the sort of model construction and analysis currently performed by experts.

The second half of this thesis examined several extensions of Gaussian processes, all of which enable the automatic determination of modeling choices that were previously set by trial and error or cross-validation. Chapter 5 characterized and visualized deep Gaussian processes, related them to existing deep neural networks, and derived novel deep kernels. Chapter 6 investigated additive GPs, a family of models consisting of sums of functions of all subsets of input variables, and showed that they have the same covariance as a GP using dropout regularization. Chapter 7 extended the GP latent variable model into a Bayesian clustering model which automatically infers the nonparametric shape of each cluster, as well as the number of clusters.

### Endnote

It seems clear that, one way or another, large parts of the existing practice of model-building will eventually be automated.

# Appendix A

## Gaussian conditionals

A standard result shows how to condition on a subset of dimensions  $\mathbf{y}_B$  of a vector  $\mathbf{y}$  having a multivariate Gaussian distribution. If

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_A \\ \mathbf{y}_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{bmatrix}\right) \quad (\text{A.1})$$

then

$$\mathbf{y}_A | \mathbf{y}_B \sim \mathcal{N}(\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} (\mathbf{x}_B - \boldsymbol{\mu}_B), \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \boldsymbol{\Sigma}_{BA}). \quad (\text{A.2})$$

This result can be used in the context of Gaussian process regression, where  $\mathbf{y}_B = [f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_N)]$  represents a set of function values observed at some subset of locations  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ , while  $\mathbf{y}_A = [f(\mathbf{x}_1^*), f(\mathbf{x}_2^*), \dots, f(\mathbf{x}_N^*)]$  represents test points whose predictive distribution we'd like to know. In this case, the necessary covariance matrices are given by:

$$\boldsymbol{\Sigma}_{AA} = k(\mathbf{X}, \mathbf{X}) \quad (\text{A.3})$$

$$\boldsymbol{\Sigma}_{AB} = k(\mathbf{X}, \mathbf{X}^*) \quad (\text{A.4})$$

$$\boldsymbol{\Sigma}_{BA} = k(\mathbf{X}^*, \mathbf{X}) \quad (\text{A.5})$$

$$\boldsymbol{\Sigma}_{BB} = k(\mathbf{X}^*, \mathbf{X}^*) \quad (\text{A.6})$$

and similarly for the mean vectors.

# Appendix B

## Kernel definitions

Here we give the formulas for all one-dimensional base kernels mentioned in the thesis. Each of these formulas is multiplied by a scale factor  $\sigma_f^2$ , which we omit for clarity.

$$C(x, x') = 1 \quad (\text{B.1})$$

$$\text{WN}(x, x') = \delta(x - x') \quad (\text{B.2})$$

$$\text{Lin}(x, x') = (x - c)(x' - c) \quad (\text{B.3})$$

$$\text{SE}(x, x') = \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \quad (\text{B.4})$$

$$\text{RQ}(x, x') = \left(1 + \frac{(x - x')^2}{2\alpha\ell^2}\right)^{-\alpha} \quad (\text{B.5})$$

$$\text{Per}(x, x') = \exp\left(-\frac{2}{\ell^2} \sin^2\left(\pi \frac{x - x'}{p}\right)\right) \quad (\text{B.6})$$

$$\cos(x, x') = \cos\left(\frac{2\pi(x - x')}{p}\right) \quad (\text{B.7})$$

$$\text{CP}(k_1, k_2)(x, x') = \sigma(x)k_1(x, x')\sigma(x') + (1 - \sigma(x))k_2(x, x')(1 - \sigma(x')) \quad (\text{B.8})$$

$$\boldsymbol{\sigma}(x, x') = \sigma(x)\sigma(x') \quad (\text{B.9})$$

$$\bar{\boldsymbol{\sigma}}(x, x') = (1 - \sigma(x))(1 - \sigma(x')) \quad (\text{B.10})$$

where  $\delta_{x,x'}$  is the Kronecker delta function,  $I_0$  is the modified Bessel function of the first kind of order zero, and other symbols are kernel parameters.  $\sigma(x) = 1/(1+\exp(-x))$ .

Equations (B.3), (B.4) and (B.11) are plotted in figure 2.1, and equations (B.2), (B.5) and (B.7) are plotted in figure 3.1. Draws from GP priors with changepoint kernels are shown in figure 2.9.

### The generalized periodic kernel

James Lloyd (personal communication) showed that the standard periodic kernel due to MacKay (1998) can be decomposed into a periodic and a constant component. He derived the equivalent periodic kernel without any constant component:

$$\text{ZPer}\emptyset(x, x') = \sigma_f^2 \frac{\exp\left(\frac{1}{\ell^2} \cos 2\pi \frac{(x-x')}{p}\right) - I_0\left(\frac{1}{\ell^2}\right)}{\exp\left(\frac{1}{\ell^2}\right) - I_0\left(\frac{1}{\ell^2}\right)} \quad (\text{B.11})$$

He further showed that its limit as the lengthscale grows is the cosine kernel:

Separating out the constant component allows us to express negative prior covariance, as well as increasing the interpretability of the resulting models. This covariance function is included in the GPML software package (Rasmussen and Nickisch, 2010), and its source can be viewed at [gaussianprocess.org/gpml/code/matlab/cov/covPeriodicNoDC.m](http://gaussianprocess.org/gpml/code/matlab/cov/covPeriodicNoDC.m).

# Appendix C

## Search operators

The model construction phase of ABCD starts with the noise kernel, WN. New kernel expressions are generated by applying search operators to the current kernel, which replace some part of the existing kernel expression with a new kernel expression.

The search used in the multidimensional regression experiments in sections 3.8.4 and 6.6 used only the following search operators:

$$\mathcal{S} \rightarrow \mathcal{S} + \mathcal{B} \tag{C.1}$$

$$\mathcal{S} \rightarrow \mathcal{S} \times \mathcal{B} \tag{C.2}$$

$$\mathcal{B} \rightarrow \mathcal{B}' \tag{C.3}$$

where  $\mathcal{S}$  represents any kernel subexpression and  $\mathcal{B}$  is any base kernel within a kernel expression. These search operators represent addition, multiplication and replacement. When the multiplication operator is applied to a subexpression which includes a sum of subexpressions, parentheses () are introduced. For instance, if rule (C.2) is applied to the subexpression  $k_1 + k_2$ , the resulting expression is  $(k_1 + k_2) \times \mathcal{B}$ .

Afterwards, we added several more search operators in order to speed up the search. These new operators do not change the set of possible models.

To accommodate changepoints and changewindows, we introduced the following additional operators to our search:

$$\mathcal{S} \rightarrow \text{CP}(\mathcal{S}, \mathcal{S}) \tag{C.4}$$

$$\mathcal{S} \rightarrow \text{CW}(\mathcal{S}, \mathcal{S}) \tag{C.5}$$

$$\mathcal{S} \rightarrow \text{CW}(\mathcal{S}, \mathcal{C}) \tag{C.6}$$

$$\mathcal{S} \rightarrow \text{CW}(\mathcal{C}, \mathcal{S}) \tag{C.7}$$

where C is the constant kernel. The last two operators result in a kernel only applying outside, or within, a certain region.

To allow the search to simplify existing expressions, we introduced the following operators:

$$\mathcal{S} \rightarrow \mathcal{B} \tag{C.8}$$

$$\mathcal{S} + \mathcal{S}' \rightarrow \mathcal{S} \tag{C.9}$$

$$\mathcal{S} \times \mathcal{S}' \rightarrow \mathcal{S} \tag{C.10}$$

where  $\mathcal{S}'$  represents any other kernel expression. We also introduced the operator

$$\mathcal{S} \rightarrow \mathcal{S} \times (\mathcal{B} + C) \tag{C.11}$$

Which allows a new base kernel to be added along with the constant kernel, for cases when multiplying by a base kernel by itself would restrict the model too much.

# Appendix D

## Example automatically generated report

The following pages of this appendix contain an entire automatically-generated report, run on a dataset measuring annual solar irradiation data from 1610 to 2011. This dataset was previously analyzed by [Lean et al. \(1995\)](#).

The structure search was run using the ABCD-interpretable variant, with base kernels SE, Lin, C, Per,  $\sigma$ , and WN.

Other example reports can be found at [mlg.eng.cam.ac.uk/Lloyd/abcdoutput/](http://mlg.eng.cam.ac.uk/Lloyd/abcdoutput/), including analyses of wheat prices, temperature records, call centre volumes, radio interference, gas production, unemployment, number of births, and wages over time.

## 1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

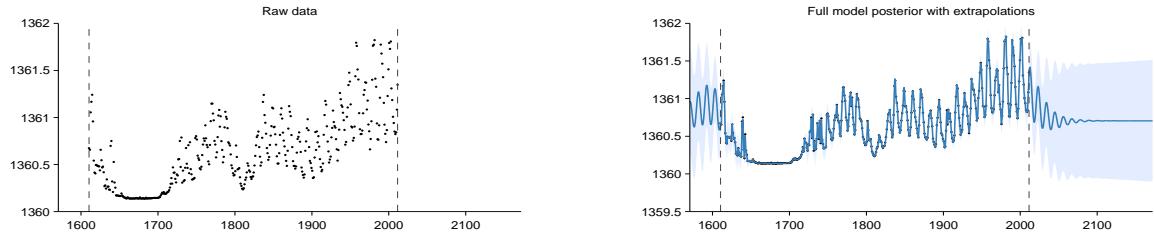


Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified nine additive components in the data. The first 4 additive components explain 92.3% of the variation in the data as shown by the coefficient of determination ( $R^2$ ) values in table 1. The first 8 additive components explain 99.2% of the variation in the data. After the first 5 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A constant.
- A constant. This function applies from 1644 until 1713.
- A smooth function. This function applies until 1644 and from 1719 onwards.
- An approximately periodic function with a period of 10.8 years. This function applies until 1644 and from 1719 onwards.
- A rapidly varying smooth function. This function applies until 1644 and from 1719 onwards.
- Uncorrelated noise.
- A rapidly varying smooth function with marginal standard deviation increasing linearly away from 1843. This function applies from 1751 onwards.
- A rapidly varying smooth function. This function applies until 1644 and from 1719 until 1751.
- A constant. This function applies from 1713 until 1719.

#	$R^2$ (%)	$\Delta R^2$ (%)	Residual $R^2$ (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	1360.65	-
1	0.0	0.0	0.0	0.33	100.0
2	35.3	35.3	35.3	0.23	29.4
3	72.5	37.2	57.5	0.18	20.7
4	92.3	19.9	72.2	0.15	16.4
5	97.8	5.5	71.4	0.15	0.4
6	97.8	0.0	0.2	0.15	0.0
7	98.4	0.5	24.8	0.15	-0.0
8	99.2	0.8	50.7	0.15	-0.0
9	100.0	0.8	100.0	0.15	-0.0

Table 1: Summary statistics for cumulative additive fits to the data. The residual coefficient of determination ( $R^2$ ) values are computed using the residuals from the previous fit as the target values; this measures how much of the residual variance is explained by each new component. The mean absolute error (MAE) is calculated using 10 fold cross validation with a contiguous block design; this measures the ability of the model to interpolate and extrapolate over moderate distances. The model is fit using the full data so the MAE values cannot be used reliably as an estimate of out-of-sample predictive performance.

## 2 Detailed discussion of additive components

### 2.1 Component 1 : A constant

This component is constant.

This component explains 0.0% of the total variance. The addition of this component reduces the cross validated MAE by 100.0% from 1360.6 to 0.3.

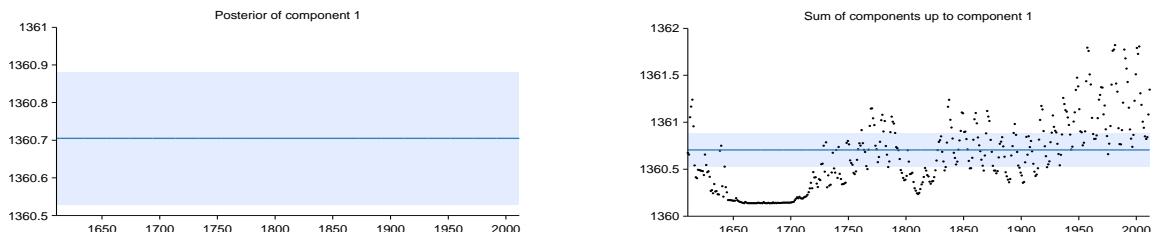


Figure 2: Posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)

## 2.2 Component 2 : A constant. This function applies from 1644 until 1713

This component is constant. This component applies from 1644 until 1713.

This component explains 35.3% of the residual variance; this increases the total variance explained from 0.0% to 35.3%. The addition of this component reduces the cross validated MAE by 29.42% from 0.33 to 0.23.

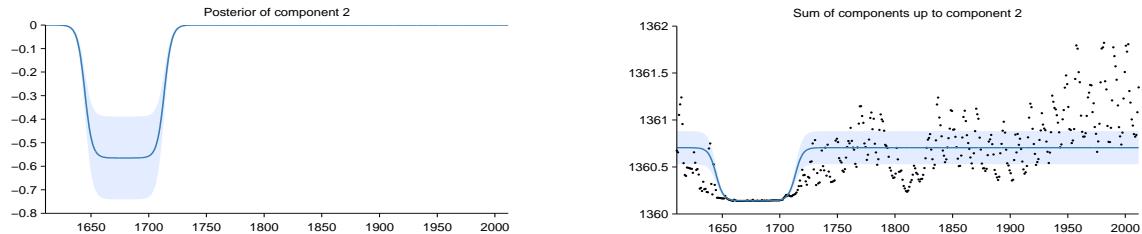


Figure 3: Posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

## 2.3 Component 3 : A smooth function. This function applies until 1644 and from 1719 onwards

This component is a smooth function with a typical lengthscale of 21.9 years. This component applies until 1644 and from 1719 onwards.

This component explains 57.5% of the residual variance; this increases the total variance explained from 35.3% to 72.5%. The addition of this component reduces the cross validated MAE by 20.66% from 0.23 to 0.18.

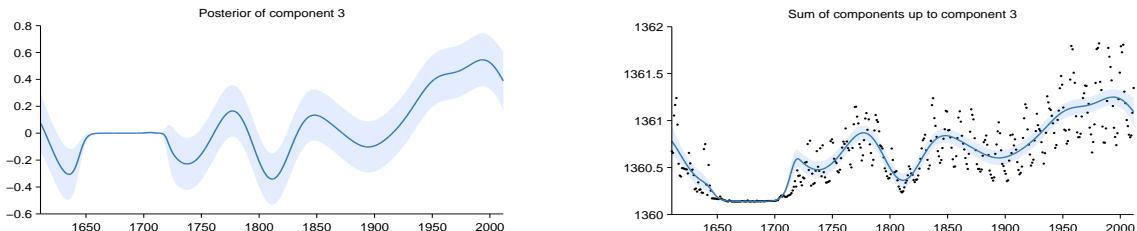


Figure 4: Posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

## 2.4 Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1644 and from 1719 onwards

This component is approximately periodic with a period of 10.8 years. Across periods the shape of the function varies smoothly with a typical lengthscale of 33.2 years. The shape of the function within each period has a typical lengthscale of 12.6 years. This component applies until 1644 and from 1719 onwards.

This component explains 72.2% of the residual variance; this increases the total variance explained from 72.5% to 92.3%. The addition of this component reduces the cross validated MAE by 16.42% from 0.18 to 0.15.

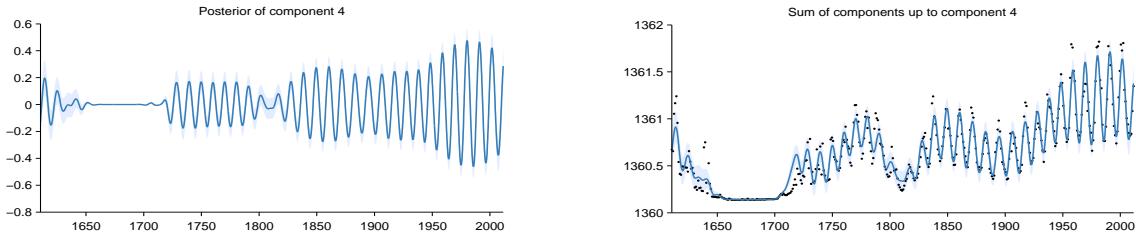


Figure 5: Posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

## 2.5 Component 5 : A rapidly varying smooth function. This function applies until 1644 and from 1719 onwards

This function is a rapidly varying but smooth function with a typical lengthscale of 1.2 years. This component applies until 1644 and from 1719 onwards.

This component explains 71.4% of the residual variance; this increases the total variance explained from 92.3% to 97.8%. The addition of this component reduces the cross validated MAE by 0.41% from 0.15 to 0.15.

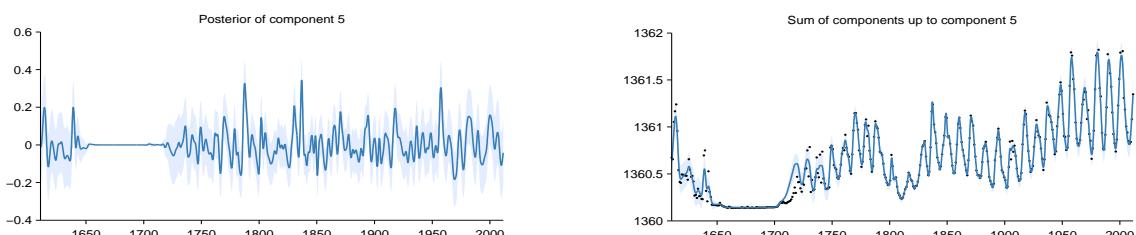


Figure 6: Posterior of component 5 (left) and the posterior of the cumulative sum of components with data (right)

## 2.6 Component 6 : Uncorrelated noise

This component models uncorrelated noise.

This component explains 0.2% of the residual variance; this increases the total variance explained from 97.8% to 97.8%. The addition of this component reduces the cross validated MAE by 0.00% from 0.15 to 0.15. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

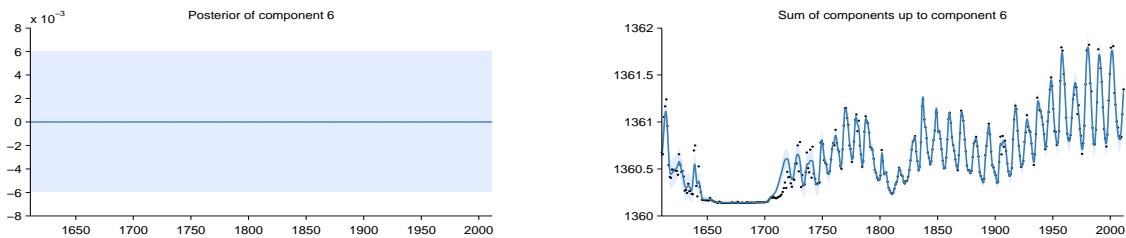


Figure 7: Posterior of component 6 (left) and the posterior of the cumulative sum of components with data (right)

## 2.7 Component 7 : A rapidly varying smooth function with marginal standard deviation increasing linearly away from 1843. This function applies from 1751 onwards

This function is a rapidly varying but smooth function with a typical lengthscale of 3.1 months. The marginal standard deviation of the function increases linearly away from 1843. This component applies from 1751 onwards.

This component explains 24.8% of the residual variance; this increases the total variance explained from 97.8% to 98.4%. The addition of this component increases the cross validated MAE by 0.00% from 0.15 to 0.15. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

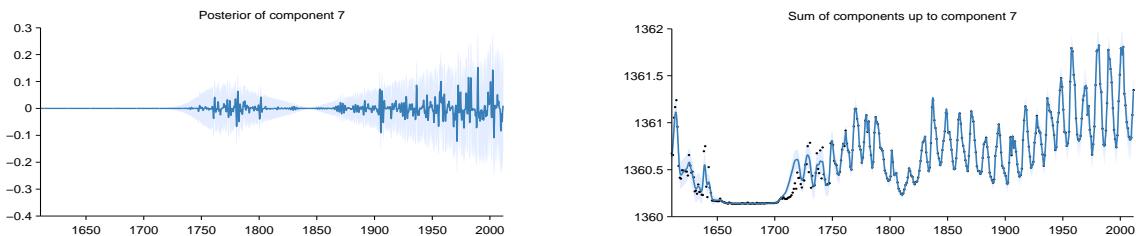


Figure 8: Posterior of component 7 (left) and the posterior of the cumulative sum of components with data (right)

## 2.8 Component 8 : A rapidly varying smooth function. This function applies until 1644 and from 1719 until 1751

This function is a rapidly varying but smooth function with a typical lengthscale of 3.1 months. This component applies until 1644 and from 1719 until 1751.

This component explains 50.7% of the residual variance; this increases the total variance explained from 98.4% to 99.2%. The addition of this component increases the cross validated MAE by 0.00% from 0.15 to 0.15. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

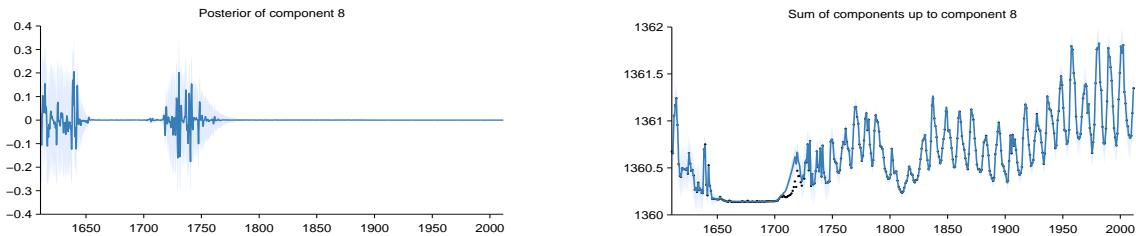


Figure 9: Posterior of component 8 (left) and the posterior of the cumulative sum of components with data (right)

## 2.9 Component 9 : A constant. This function applies from 1713 until 1719

This component is constant. This component applies from 1713 until 1719.

This component explains 100.0% of the residual variance; this increases the total variance explained from 99.2% to 100.0%. The addition of this component increases the cross validated MAE by 0.01% from 0.15 to 0.15. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

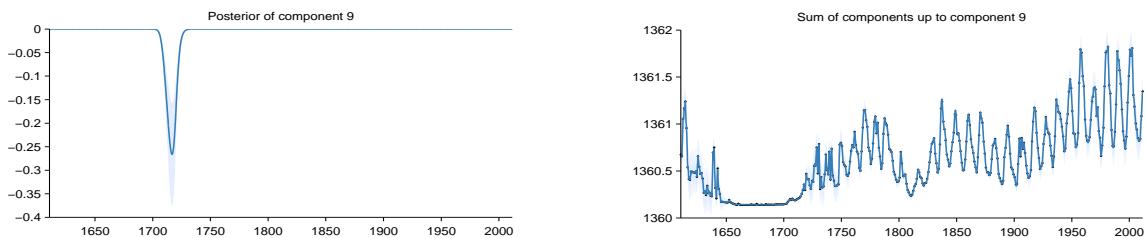


Figure 10: Posterior of component 9 (left) and the posterior of the cumulative sum of components with data (right)

### 3 Extrapolation

Summaries of the posterior distribution of the full model are shown in figure 11. The plot on the left displays the mean of the posterior together with pointwise variance. The plot on the right displays three random samples from the posterior.



Figure 11: Full model posterior. Mean and pointwise variance (left) and three random samples (right)

## Appendix E

# Inference in the warped mixture model

### Detailed definition of model

The iWMM assumes that the latent density is an infinite mixture of Gaussians:

$$p(\mathbf{x}) = \sum_{c=1}^{\infty} \lambda_c \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_c, \mathbf{R}_c^{-1}) \quad (\text{E.1})$$

where  $\lambda_c$ ,  $\boldsymbol{\mu}_c$  and  $\mathbf{R}_c$  is the mixture weight, mean, and precision matrix of the  $c^{\text{th}}$  mixture component. We place a conjugate Gaussian-Wishart priors on the Gaussian parameters  $\{\boldsymbol{\mu}_c, \mathbf{R}_c\}$ :

$$p(\boldsymbol{\mu}_c, \mathbf{R}_c) = \mathcal{N}(\boldsymbol{\mu}_c | \mathbf{u}, (r \mathbf{R}_c)^{-1}) \mathcal{W}(\mathbf{R}_c | \mathbf{S}^{-1}, \nu), \quad (\text{E.2})$$

where  $\mathbf{u}$  is the mean of  $\boldsymbol{\mu}_c$ ,  $r$  is the relative precision of  $\boldsymbol{\mu}_c$ ,  $\mathbf{S}^{-1}$  is the scale matrix for  $\mathbf{R}_c$ , and  $\nu$  is the number of degrees of freedom for  $\mathbf{R}_c$ . The Wishart distribution is defined as:

$$\mathcal{W}(\mathbf{R} | \mathbf{S}^{-1}, \nu) = \frac{1}{G} |\mathbf{R}|^{\frac{\nu-Q-1}{2}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{S}\mathbf{R})\right), \quad (\text{E.3})$$

where  $G$  is the normalizing constant.

Because we use conjugate Gaussian-Wishart priors for the parameters of the Gaussian mixture components, we can analytically integrate out those parameters given the assignments of points to components. Let  $z_n$  be the assignment of the  $n^{\text{th}}$  point. The prior probability of latent coordinates  $\mathbf{X}$  given latent cluster assignments  $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N)$

factorizes over clusters, and can be obtained in closed-form by integrating out the Gaussian parameters  $\{\boldsymbol{\mu}_c, \mathbf{R}_c\}$  to give:

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{S}, \nu, r) = \prod_{c=1}^{\infty} \pi^{-\frac{N_c Q}{2}} \frac{r^{Q/2} |\mathbf{S}|^{\nu/2}}{r_c^{Q/2} |\mathbf{S}_c|^{\nu_c/2}} \times \prod_{q=1}^Q \frac{\Gamma\left(\frac{\nu_c+1-q}{2}\right)}{\Gamma\left(\frac{\nu+1-q}{2}\right)}, \quad (\text{E.4})$$

where  $N_c$  is the number of data points assigned to the  $c^{\text{th}}$  component,  $\Gamma(\cdot)$  is the Gamma function, and

$$r_c = r + N_c, \quad \nu_c = \nu + N_c, \quad \mathbf{u}_c = \frac{r\mathbf{u} + \sum_{n:z_n=c} \mathbf{x}_n}{r + N_c}, \quad (\text{E.5})$$

$$\text{and} \quad \mathbf{S}_c = \mathbf{S} + \sum_{n:z_n=c} \mathbf{x}_n \mathbf{x}_n^\top + r\mathbf{u}\mathbf{u}^\top - r_c \mathbf{u}_c \mathbf{u}_c^\top, \quad (\text{E.6})$$

are the posterior Gaussian-Wishart parameters of the  $c^{\text{th}}$  component (Murphy, 2007).

To model the cluster assignments, we use a Dirichlet process (MacEachern and Müller, 1998) with concentration parameter  $\eta$ . Under a Dirichlet process prior, the probability of observing a particular cluster assignment matrix  $\mathbf{Z}$  depends only on the partition induced, and is given by the Chinese restaurant process:

$$p(\mathbf{Z}|\eta) = \frac{\Gamma(\eta)\eta^C}{\Gamma(\eta+N)} \prod_{c=1}^C \Gamma(N_c) \quad (\text{E.7})$$

where  $C$  is the number of components for which  $N_c > 0$ , and  $N$  is the total number of datapoints.

The joint distribution of observed coordinates, latent coordinates, and cluster assignments is given by

$$p(\mathbf{Y}, \mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \mathbf{S}, \nu, \mathbf{u}, r, \eta) = p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{X}|\mathbf{Z}, \mathbf{S}, \nu, \mathbf{u}, r)p(\mathbf{Z}|\eta), \quad (\text{E.8})$$

where the factors in the right hand side can be calculated by equations (7.5), (E.4) and (E.7), respectively.

## Details of inference

After analytically integrating out the parameters of the Gaussian mixture components, the only remaining variables to infer are the latent points  $\mathbf{X}$ , the cluster assignments  $\mathbf{Z}$ , and the kernel parameters  $\boldsymbol{\theta}$ . We'll estimate the posterior over these parameters using Markov chain Monte Carlo. In particular, we'll alternate between collapsed Gibbs

sampling of each row of  $\mathbf{Z}$ , and Hamiltonian Monte Carlo sampling of  $\mathbf{X}$  and  $\boldsymbol{\theta}$ .

First, we explain collapsed Gibbs sampling for the cluster assignments  $\mathbf{Z}$ . Given a sample of  $\mathbf{X}$ ,  $p(\mathbf{Z}|\mathbf{X}, \mathbf{S}, \nu, \mathbf{u}, r, \eta)$  does not depend on  $\mathbf{Y}$ . This lets us resample cluster assignments, integrating out the iGMM likelihood in closed form. Given the current state of all but one latent component  $z_n$ , a new value for  $z_n$  is sampled with the following probability:

$$p(z_n = c|\mathbf{X}, \mathbf{Z}_{\setminus n}, \mathbf{S}, \nu, \mathbf{u}, r, \eta) \propto \begin{cases} N_{c \setminus n} \cdot p(\mathbf{x}_n|\mathbf{X}_{c \setminus n}, \mathbf{S}, \nu, \mathbf{u}, r) & \text{existing components} \\ \eta \cdot p(\mathbf{x}_n|\mathbf{S}, \nu, \mathbf{u}, r) & \text{a new component} \end{cases} \quad (\text{E.9})$$

where  $\mathbf{X}_c = \{\mathbf{x}_n | z_n = c\}$  is the set of latent coordinates assigned to the  $c^{\text{th}}$  component, and  $\setminus n$  represents the value or set when excluding the  $n^{\text{th}}$  data point. We can analytically calculate  $p(\mathbf{x}_n|\mathbf{X}_{c \setminus n}, \mathbf{S}, \nu, \mathbf{u}, r)$  as follows:

$$p(\mathbf{x}_n|\mathbf{X}_{c \setminus n}, \mathbf{S}, \nu, \mathbf{u}, r) = \pi^{-\frac{N_{c \setminus n} Q}{2}} \frac{r_{c \setminus n}^{Q/2} |\mathbf{S}_{c \setminus n}|^{\nu_{c \setminus n}/2}}{r'_{c \setminus n}^{Q/2} |\mathbf{S}'_{c \setminus n}|^{\nu'_{c \setminus n}/2}} \times \prod_{d=1}^Q \frac{\Gamma\left(\frac{\nu'_{c \setminus n} + 1 - d}{2}\right)}{\Gamma\left(\frac{\nu_{c \setminus n} + 1 - d}{2}\right)}, \quad (\text{E.10})$$

where  $r'_c$ ,  $\nu'_c$ ,  $\mathbf{u}'_c$  and  $\mathbf{S}'_c$  represent the posterior on Gaussian-Wishart parameters of the  $c^{\text{th}}$  component when the  $n^{\text{th}}$  data point has been assigned to it. We efficiently calculate the determinant by using the rank-one Cholesky update. A special case of equation (E.10) gives the likelihood for a new component:  $p(\mathbf{x}_n|\mathbf{S}, \nu, \mathbf{u}, r)$ .

## Gradients for Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) sampling of  $\mathbf{X}$  from posterior  $p(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}, \mathbf{S}, \nu, \mathbf{u}, r)$ , requires computing the gradient of the log-unnormalized-posterior with respect to  $\mathbf{X}$ :

$$\frac{\partial}{\partial \mathbf{X}} \left[ \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) + \log p(\mathbf{X}|\mathbf{Z}, \mathbf{S}, \nu, \mathbf{u}, r) \right] \quad (\text{E.11})$$

The first term of gradient (E.11) can be calculated by

$$\frac{\partial \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})}{\partial \mathbf{X}} = \frac{\partial \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})}{\partial \mathbf{K}} \frac{\partial \mathbf{K}}{\partial \mathbf{X}} = \left[ -\frac{1}{2} D \mathbf{K}^{-1} + \frac{1}{2} \mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^T \mathbf{K}^{-1} \right] \left[ \frac{\partial \mathbf{K}}{\partial \mathbf{X}} \right], \quad (\text{E.12})$$

where for an SE + WN kernel with the same lengthscale  $\ell$  on all dimensions,

$$\frac{\partial k(\mathbf{x}_n, \mathbf{x}_m)}{\partial \mathbf{x}_n} = -\frac{\sigma_f^2}{\ell^2} \exp\left(-\frac{1}{2\ell^2}(\mathbf{x}_n - \mathbf{x}_m)^\top(\mathbf{x}_n - \mathbf{x}_m)\right)(\mathbf{x}_n - \mathbf{x}_m). \quad (\text{E.13})$$

The second term of (E.11) is

$$\frac{\partial \log p(\mathbf{X}|\mathbf{Z}, \mathbf{S}, \nu, \mathbf{u}, r)}{\partial \mathbf{x}_n} = -\nu_{z_n} \mathbf{S}_{z_n}^{-1}(\mathbf{x}_n - \mathbf{u}_{z_n}). \quad (\text{E.14})$$

We also infer kernel parameters  $\boldsymbol{\theta}$  via HMC, using the gradient of the log unnormalized posterior with respect to the kernel parameters and an improper uniform prior.

### Posterior predictive density

In the GP-LVM, the predictive density of at test point  $\mathbf{y}_*$  is usually computed by finding the point  $\mathbf{x}_*$  which has the highest probability of being mapped to  $\mathbf{y}_*$ , then using the density of  $p(\mathbf{x}_*)$  and the Jacobian of the warping at that point to approximate the density at  $\mathbf{y}_*$ . When inference is done this way, approximating the predictive density only requires solving a single optimization for each  $\mathbf{y}_*$ .

For our model, we use approximate integration to estimate  $p(\mathbf{y}_*)$ . This is done for two reasons: First, multiple latent points (possibly from different clusters) can map to the same observed point, meaning the standard method can underestimate  $p(\mathbf{y}_*)$ . Second, because we do not optimize the latent coordinates of training points, but instead sample them, we would need to optimize each  $p(\mathbf{x}_*)$  separately for each sample in the Markov chain. One advantage of our method is that it gives estimates for all  $p(\mathbf{y}_*)$  at once, but it may not be accurate in very high dimensions.

The posterior density in the observed space given the training data is

$$\begin{aligned} p(\mathbf{y}_*|\mathbf{Y}) &= \iint p(\mathbf{y}_*, \mathbf{x}_*, \mathbf{X}|\mathbf{Y}) d\mathbf{x}_* d\mathbf{X} \\ &= \iint p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{Y}) p(\mathbf{x}_*|\mathbf{X}, \mathbf{Y}) p(\mathbf{X}|\mathbf{Y}) d\mathbf{x}_* d\mathbf{X}. \end{aligned} \quad (\text{E.15})$$

We first approximate  $p(\mathbf{X}|\mathbf{Y})$  using samples from the Gibbs and Hamiltonian Monte Carlo chains. We then approximate  $p(\mathbf{x}_*|\mathbf{X}, \mathbf{Y})$  by sampling points from the posterior density in the latent space and warping them, using the following procedure:

1. Draw latent cluster assignment  $z_* \sim \text{Mult}(\frac{N_1}{N+\eta}, \dots, \frac{N_C}{N+\eta}, \frac{\eta}{N+\eta})$
2. Draw latent cluster precision matrix  $\mathbf{R}_* \sim \mathcal{W}(\mathbf{S}_{z_*}^{-1}, \nu_{z_*})$

- 
3. Draw latent cluster mean  $\boldsymbol{\mu}_\star \sim \mathcal{N}(\mathbf{u}_{z_\star}, (r_{z_\star} \mathbf{R}_\star)^{-1})$
  4. Draw latent coordinates  $\mathbf{x}_\star \sim \mathcal{N}(\boldsymbol{\mu}_\star, \mathbf{R}_\star^{-1})$
  5. Draw observed coordinates  $\mathbf{y}_\star \sim \mathcal{N}(\mathbf{k}_\star^\top \mathbf{K}^{-1} \mathbf{Y}, k(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{k}_\star^\top \mathbf{K}^{-1} \mathbf{k}_\star)$

If  $z_\star$  is assigned to a new component in step 1, the prior Gaussian-Wishart distribution (E.2) is used for sampling in steps 2 and 3. The density drawn from in step 5 is simply the predictive distribution of a GP, where  $\mathbf{k}_\star = (k(\mathbf{x}_\star, \mathbf{x}_1), \dots, k(\mathbf{x}_\star, \mathbf{x}_N))^\top$ .

Each step of this sampling procedure draws from the exact conditional distribution, so the Monte Carlo estimate of the predictive density  $p(\mathbf{y}_\star | \mathbf{X}, \mathbf{Y})$  will converge to the true marginal distribution as the number of samples increases. Since the observations  $\mathbf{y}_\star$  are conditionally normally distributed, each one adds a smooth contribution to the empirical Monte Carlo estimate of the posterior density, as opposed to a collection of point masses.

## Source code

A reference implementation of the above algorithms is available at [github.com/duvenaud/warped-mixtures](https://github.com/duvenaud/warped-mixtures).

# References

- Ryan P. Adams and Zoubin Ghahramani. Archipelago: Nonparametric Bayesian semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1–8. ACM, 2009. (pages 103 and 106)
- Ryan P. Adams, Hanna M. Wallach, and Zoubin Ghahramani. Learning the structure of deep sparse graphical models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010. (pages 76 and 140)
- Mark A. Armstrong, Gérard Iooss, and Daniel D. Joseph. *Groups and symmetry*. Springer, 1988. (page 23)
- Francis R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, pages 105–112, 2009a. (page 41)
- Francis R. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. *arXiv preprint arXiv:0909.0844*, 2009b. (pages 87, 91, and 93)
- Francis R. Bach, Gert R.G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004. (pages 41 and 43)
- Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>. (page 91)
- Pierre Baldi and Peter J. Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems*, pages 2814–2822, 2013. (page 85)
- Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey M. Siskind, Jarrell Waggoner, Song Wang, Jinlian Wei, Yifan Yin, and Zhiqi Zhang. Video in sentences out. In *Conference on Uncertainty in Artificial Intelligence*, 2012. (page 56)
- Richard Bellman. Dynamic programming and Lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767, 1956. (page 12)
- Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166, 1994. (page 77)

- Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. The curse of highly variable functions for local kernel machines. *Advances in Neural Information Processing Systems*, 18:107–114, 2006. ISSN 1049-5258. (pages 71 and 84)
- Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013. (page 140)
- Wu Bing, Zhang Wen-qiong, Chen Ling, and Liang Jia-hong. A GP-based kernel construction and optimization method for RVM. In *International Conference on Computer and Automation Engineering (ICCAE)*, volume 4, pages 419–423, 2010. (page 42)
- Salomon Bochner. *Lectures on Fourier integrals*, volume 42. Princeton University Press, 1959. (pages 21 and 41)
- George E.P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time series analysis: forecasting and control*. John Wiley & Sons, 1970. (pages 38 and 43)
- Wenbo Cao and Robert Haralick. Nonlinear manifold clustering by dimensionality. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 920–924. IEEE, 2006. (page 101)
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. (page 106)
- Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pages 342–350, 2009. (pages 71 and 77)
- Corinna Cortes and Vladimir N. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. (page 16)
- Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013. (pages 58, 76, 101, 139, and 140)
- Eyal Dechter, Jon Malmaud, Ryan P. Adams, and Joshua B. Tenenbaum. Bootstrap learning via modular concept discovery. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1302–1309. AAAI Press, 2013. (page 42)
- Laura Diosan, Alexandrina Rogozan, and Jean-Pierre Pecuchet. Evolving kernel functions for SVMs by genetic programming. In *Machine Learning and Applications, 2007*, pages 19–24. IEEE, 2007. (page 42)
- Nicolas Durrande, David Ginsbourger, and Olivier Roustant. Additive kernels for Gaussian process modeling. *arXiv preprint arXiv:1103.4023*, 2011. (page 87)
- David Duvenaud, Hannes Nickisch, and Carl E. Rasmussen. Additive Gaussian processes. In *Advances in Neural Information Processing Systems 24*, pages 226–234, Granada, Spain, 2011. (page 79)

- David Duvenaud, James Robert Lloyd, Roger B. Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, 2013. (pages 8 and 30)
- David Duvenaud, Oren Rippel, Ryan P. Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. In *17th International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland, April 2014. (page 59)
- Daniel Eaton and Kevin Murphy. Bayesian structure learning using dynamic programming and MCMC. In *Conference on Uncertainty in Artificial Intelligence*, 2007. (page 31)
- Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *Advances in Neural Information Processing Systems*, pages 55–63, 2011. (page 101)
- Emily B. Fox and David B. Dunson. Multiresolution Gaussian processes. In *Advances in Neural Information Processing Systems 25*. MIT Press, 2013. (page 43)
- Nir Friedman and Daphne Koller. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–126, 2003. (page 31)
- M. Ganeshalingam and Timonthy W. Gowers. A fully automatic problem solver with human-style output. *arXiv preprint arXiv:1309.4501*, 2013. (page 56)
- Roman Garnett, Michael A. Osborne, Steven Reece, Alex Rogers, and Stephen J. Roberts. Sequential Bayesian prediction in the presence of changepoints and faults. *The Computer Journal*, 53(9):1430–1446, 2010. (pages 33 and 43)
- Andreas Geiger, Raquel Urtasun, and Trevor Darrell. Rank priors for continuous non-linear dimensionality reduction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 880–887. IEEE, 2009. (page 100)
- Andrew Gelman. Why waste time philosophizing?, 2013. URL <http://andrewgelman.com/2013/02/11/why-waste-time-philosophizing/>. (page 31)
- Andrew Gelman and Cosma R. Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 2012. (page 31)
- Robert Gens and Pedro Domingos. Learning the structure of sum-product networks. In *International Conference on Machine learning*, 2013. (page 78)
- Zoubin Ghahramani and M.J. Beal. Variational inference for Bayesian mixtures of factor analysers. *Advances in Neural Information Processing Systems*, 12:449–455, 2000. (page 100)
- Elad Gilboa, Yunus Saatçi, and John Cunningham. Scaling multidimensional inference for structured Gaussian processes. In *Proceedings of the 30th International Conference on Machine Learning*, 2013. (page 87)

- David Ginsbourger, Xavier Bay, Olivier Roustant, and Laurent Carraro. Argumentwise invariant kernels for the approximation of invariant functions. In *Annales de la Faculté de Sciences de Toulouse*, 2012. (page 23)
- David Ginsbourger, Olivier Roustant, and Nicolas Durrande. Invariances of random fields paths, with applications in Gaussian process regression. *arXiv preprint arXiv:1308.1359 [math.ST]*, August 2013. (pages 23 and 24)
- Mehmet Gönen and Ethem Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011. (pages 33, 41, and 43)
- Noah D. Goodman, Vikash K. Mansinghka, Daniel M. Roy, K. Bonawitz, and Joshua B. Tenenbaum. Church: A language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 220–229, 2008. (page 140)
- Noah D. Goodman, Tomer D. Ullman, and Joshua B. Tenenbaum. Learning a theory of causality. *Psychological review*, 118(1):110, 2011. (page 13)
- Daniel B Graham and Nigel M Allinson. Characterizing virtual eigensignatures for general purpose face recognition. *Face Recognition: From Theory to Applications*, 163:446–456, 1998. (page 103)
- Roger B. Grosse. *Model Selection in Compositional Spaces*. PhD thesis, Massachusetts Institute of Technology, 2014. (page 140)
- Roger B. Grosse, Ruslan Salakhutdinov, William T. Freeman, and Joshua B. Tenenbaum. Exploiting compositionality to explore a large space of model structures. In *Uncertainty in Artificial Intelligence*, 2012. (pages 31 and 42)
- Chong Gu. *Smoothing spline ANOVA models*. Springer Verlag, 2002. ISBN 0387953531. (page 40)
- Trevor J. Hastie and Robert J. Tibshirani. *Generalized additive models*. Chapman & Hall/CRC, 1990. (pages 16, 40, and 80)
- James Hensman, Nicolo Fusi, and Neil D. Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013. (page 5)
- James Hensman, Andreas Damianou, and Neil D. Lawrence. Deep Gaussian processes for large datasets. In *Artificial Intelligence and Statistics Late-breaking Posters*, 2014a. (page 58)
- James Hensman, Nicolo Fusi, Ricardo Andrade, Nicolas Durrande, Alan Saul, Max Zwiessele, and Neil D. Lawrence. GPy: A Gaussian process framework in python, 2014b. <https://github.com/SheffieldML/GPy>. (page 5)
- Michiel Hermans and Benjamin Schrauwen. Recurrent kernel machines: Computing with infinite echo state networks. *Neural Computation*, 24(1):104–133, 2012. (page 77)
- Geoffrey Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. (page 85)

- Marcus Hutter. The fastest and shortest algorithm for all well-defined problems. *International Journal of Foundations of Computer Science*, 13(03):431–443, 2002. (page 140)
- Rob J. Hyndman. Time series data library, accessed July 2013. URL <http://data.is/TSDLdemo>. (page 43)
- Tomoharu Iwata, David Duvenaud, and Zoubin Ghahramani. Warped mixtures for non-parametric cluster shapes. In *29th Conference on Uncertainty in Artificial Intelligence*, Bellevue, Washington, July 2013. (page 95)
- Edwin T. Jaynes. Highly informative priors. In *Proceedings of the Second International Meeting on Bayesian Statistics*, 1985. (page 30)
- Cari G. Kaufman and Stephan R. Sain. Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–150, 2010. (page 87)
- Charles Kemp and Joshua B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008. (page 41)
- Edward D. Klenske, Melanie N. Zeilinger, Bernhard Schölkopf, and Philipp Hennig. Non-parametric dynamics estimation for time periodic systems. In *51st Annual Allerton Conference on Communication, Control, and Computing*, pages 486–493, Oct 2013. (page 40)
- Imre Risi Kondor. *Group theoretical methods in machine learning*. PhD thesis, Columbia University, 2008. (pages 23 and 26)
- Gabriel Kronberger and Michael Kommenda. Evolution of covariance functions for Gaussian process regression using genetic programming. *arXiv preprint arXiv:1305.3794*, 2013. (page 42)
- Neil D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16:329–336, 2004. (pages 96 and 100)
- Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005. (page 27)
- Neil D. Lawrence. Local distance preservation in the GP-LVM through back constraints. In *Proceedings of the 23rd international conference on Machine learning*, pages 513–520. ACM Press, 2006. (page 107)
- Neil D. Lawrence and Andrew J. Moore. Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, pages 481–488. ACM, 2007. (page 76)
- Neil D. Lawrence and Raquel Urtasun. Non-linear matrix factorization with Gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 601–608. ACM, 2009. (page 100)

- Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl E. Rasmussen, and Aníbal R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 99:1865–1881, 2010. (pages 33 and 41)
- Judith Lean, Juerg Beer, and Raymond Bradley. Reconstruction of solar irradiance since 1610: Implications for climate change. *Geophysical Research Letters*, 22(23): 3195–3198, 1995. (pages 52 and 147)
- Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361, 1995. (page 26)
- Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area v2. In *Advances in Neural Information Processing Systems*, pages 873–880, 2007. (page 78)
- Doug Lerner and Dan Asimov. The Sudanese Möbius band. In *SIGGRAPH Electronic Theatre*, 1984. (page 28)
- Leonid A. Levin. Universal sequential search problems. *Problemy Peredachi Informatsii*, 9(3):115–116, 1973. (page 140)
- Ming Li and Paul M.B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 1997. (page 140)
- Percy Liang, Michael I. Jordan, and Dan Klein. Learning programs: A hierarchical Bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 639–646, 2010. (pages 42 and 140)
- Douglas A. Lind, William G. Marchal, and Samuel Adam Wathen. *Basic statistics for business and economics*. McGraw-Hill/Irwin Boston, 2006. (pages 33 and 43)
- James Robert Lloyd. GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *International Journal of Forecasting*, 2013. (page 40)
- James Robert Lloyd, David Duvenaud, Roger B. Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2014. (pages 30, 41, 47, and 56)
- Ian G. Macdonald. *Symmetric functions and Hall polynomials*. Oxford University Press, USA, 1998. ISBN 0198504500. (page 83)
- Steven N. MacEachern and Peter Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, pages 223–238, 1998. (page 156)
- David J.C. MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992. (page 3)
- David J.C. MacKay. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998. (pages 25 and 144)

- David J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University press, 2003. (pages 3 and 99)
- Vikash Mansinghka, Daniel Selsam, and Yura Perov. Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv preprint arXiv:1404.0099*, 2014. (page 140)
- James Martens. Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, pages 735–742, 2010. (page 78)
- John H. R. Maunsell and David C. van Essen. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of neuroscience*, 3(12):2563–2586, 1983. (page 70)
- James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pages 415–446, 1909. (page 21)
- Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006. (page 12)
- Jeffrey W. Miller and Matthew T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems 26*, pages 199–206. Curran Associates, Inc., 2013. (page 108)
- Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer’s theorem, feature maps, and smoothing. In *Learning theory*, pages 154–168. Springer, 2006. (page 21)
- Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, volume 17, pages 362–369, 2001. (page 90)
- @ML\_Hipster. “...essentially, all models are wrong, but yours are stupid too.” – G.E.P. Box in a less than magnanimous mood., 2013. URL [https://twitter.com/ML\\_Hipster/status/394577463990181888](https://twitter.com/ML_Hipster/status/394577463990181888). (page 1)
- Grégoire Montavon, Mikio L. Braun, and Klaus-Robert Müller. Layer-wise analysis of deep networks with Gaussian kernels. *Advances in Neural Information Processing Systems*, 23:1678–1686, 2010. (page 77)
- Kevin P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, Computer Science Department, University of British Columbia, 2007. (page 156)
- Radford M. Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995. (pages 16, 59, 70, 76, and 77)
- Radford M. Neal. Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629, 2003. (page 108)
- John Ashworth Nelder and Robert W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. (page 80)

- Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002.  
(page 101)
- Hannes Nickisch and Carl E. Rasmussen. Gaussian mixture modeling with Gaussian process latent variable models. *Pattern Recognition*, pages 272–282, 2010. (page 100)
- Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.  
(page 90)
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *arXiv preprint arXiv:1211.5063*, 2012.  
(page 77)
- Tony A. Plate. Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models. *Behaviormetrika*, 26:29–50, 1999. ISSN 0385-7417.  
(pages 33, 40, and 87)
- Daniel Preotiuc-Pietro and Trevor Cohn. A temporal model of text periodicities using Gaussian processes. In *Conference on Empirical Methods on Natural Language Processing*, pages 977–988. ACL, 2013.  
(page 40)
- Joaquin Quiñonero-Candela and Carl E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.  
(pages 5, 99, and 108)
- Herschel Rabitz and Ömer F. Aliş. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25(2-3):197–233, 1999.  
(page 88)
- William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, pages 846–850, 1971.  
(page 106)
- Carl E. Rasmussen. The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 12(5.2):2, 2000.  
(pages 98 and 127)
- Carl E. Rasmussen and Zoubin Ghahramani. Occam’s razor. *Advances in Neural Information Processing Systems*, pages 294–300, 2001.  
(pages 3 and 36)
- Carl E. Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, December 2010.  
(pages 5 and 144)
- Carl E. Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*, volume 38. The MIT Press, Cambridge, MA, USA, 2006.  
(pages 1, 5, 37, 39, and 123)
- Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller, Yoshua Bengio, Yann Dauphin, and Xavier Glorot. Higher order contractive auto-encoder. In *Machine Learning and Knowledge Discovery in Databases*, pages 645–660. Springer, 2011a.  
(page 66)

- Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, pages 833–840, 2011b. (pages 66 and 67)
- Oren Rippel and Ryan P. Adams. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*, 2013. (pages 76 and 140)
- Carlos E. Rodríguez and Stephen G. Walker. Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Statistics and Computing*, pages 1–15, 2012. (page 101)
- David Ruppert, Matthew P. Wand, and Raymond J. Carroll. *Semiparametric regression*, volume 12. Cambridge University Press, 2003. (pages 33 and 40)
- Yunus Saatçi, Ryan D. Turner, and Carl E. Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934, 2010. (page 43)
- Ruslan Salakhutdinov and Geoffrey Hinton. Using deep belief nets to learn covariance kernels for Gaussian processes. *Advances in Neural information processing systems*, 20:1249–1256, 2008. (page 41)
- Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009. (page 140)
- Mathieu Salzmann, Raquel Urtasun, and Pascal Fua. Local deformation models for monocular 3D shape recovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. (page 100)
- Andrew Saxe, Pang W. Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y. Ng. On random weights and unsupervised feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1089–1096, 2011. (page 78)
- Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Dynamics of learning in deep linear neural networks. In *NIPS Workshop on Deep Learning*, 2013. (page 78)
- Jürgen Schmidhuber. The speed prior: a new simplicity measure yielding near-optimal computable predictions. In *Computational Learning Theory*, pages 216–228. Springer, 2002. (page 140)
- Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. ISSN 1095-9203. doi: 10.1126/science.1165893. (page 41)
- Michael Schmidt and Hod Lipson. Eureqa [software], 2013. URL <http://www.eureqa.com>. (pages 41 and 43)
- Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. (page 36)

- Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 2006.  
(pages 5, 99, and 108)
- Ercan Solak, Roderick Murray-Smith, William E. Leithead, Douglas J. Leith, and Carl E. Rasmussen. Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems 16*, 2003.  
(page 65)
- Ray J. Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22, 1964.  
(page 140)
- Nitish Srivastava. Improving neural networks with dropout. Master’s thesis, University of Toronto, 2013.  
(page 85)
- Christian Steinruecken. *Lossless Data Compression*. PhD thesis, Cavendish Laboratory, University of Cambridge, 2014.  
(pages 42 and 140)
- Mark O. Stitson, Alex Gammerman, Vladimir Vapnik, Volodya Vovk, Chris Watkins, and Jason Weston. Support vector regression with ANOVA decomposition kernels. *Advances in kernel methods: Support vector learning*, pages 285–292, 1999.  
(page 88)
- Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.  
(page 108)
- Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.  
(page 100)
- Michalis Titsias and Neil D. Lawrence. Bayesian Gaussian process latent variable model. *International Conference on Artificial Intelligence and Statistics*, 2010.  
(page 100)
- Ljupčo Todorovski and Sašo Džeroski. Declarative bias in equation discovery. In *International Conference on Machine Learning*, pages 376–384, 1997.  
(page 41)
- Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14(1):1175–1179, 2013.  
(page 5)
- Vladimir N. Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998.  
(page 88)
- Grace Wahba. *Spline models for observational data*. Society for Industrial Mathematics, 1990. ISBN 0898712440.  
(pages 40 and 88)
- Chunyi Wang and Radford M. Neal. Gaussian process regression with heteroscedastic or non-gaussian residuals. *arXiv preprint arXiv:1212.6246*, 2012.  
(page 76)
- Jingdong Wang, Jianguo Lee, and Changshui Zhang. Kernel trick embedded Gaussian mixture model. In *Algorithmic Learning Theory*, pages 159–174. Springer, 2003.  
(page 101)

- Sida Wang and Christopher Manning. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 118–126, 2013.  
(page 85)
- Takashi Washio, Hiroshi Motoda, and Yuji Niwa. Discovering admissible model equations from observed data based on scale-types and identity constraints. In *International Joint Conference On Artificial Intelligence*, volume 16, pages 772–779, 1999.  
(page 41)
- Wikimedia Commons. Stereographic projection of a Sudanese Möbius band, 2005. URL <http://commons.wikimedia.org/wiki/File:MobiusSnail2B.png>.  
(page 28)
- Andrew G. Wilson and Ryan P. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1067–1075, 2013.  
(pages 33, 41, and 43)
- Andrew G. Wilson, David A. Knowles, and Zoubin Ghahramani. Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning*, pages 599–606, 2012.  
(page 76)
- I-Cheng Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.  
(page 16)
- Yichuan Zhang and Charles A. Sutton. Quasi-Newton methods for Markov chain Monte Carlo. *Advances in Neural Information Processing Systems*, pages 2393–2401, 2011.  
(page 108)