Adaptive Multiple Importance Sampling for Gaussian Processes and Its Application in Social Signal Processing

Xiaoyu Xiong

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Computing Science

College of Science and Engineering
University of Glasgow

June 2017

ABSTRACT

Social signal processing aims to automatically understand and interpret social signals (e.g. facial expressions and prosody) generated during human-human and human-machine interactions. Automatic interpretation of social signals involves two fundamentally important aspects: feature extraction and machine learning. So far, machine learning approaches applied to social signal processing have mainly focused on parametric approaches (e.g. linear regression) or non-parametric models such as support vector machine (SVM). However, these approaches fall short of taking into account any uncertainty as a result of model misspecification or lack interpretability for analyses of scenarios in social signal processing. Consequently, they are less able to understand and interpret human behaviours effectively.

Gaussian processes (GPs), that have gained popularity in data analysis, offer a solution to these limitations through their attractive properties: being non-parametric enables them to flexibly model data and being probabilistic makes them capable of quantifying uncertainty. In addition, a proper parametrisation in the covariance function makes it possible to gain insights into the application under study.

However, these appealing properties of GP models hinge on an accurate characterisation of the posterior distribution with respect to the covariance parameters. This is normally done by means of standard Markov chain Monte Carlo (MCMC) algorithms, which require repeated expensive calculations involving the marginal likelihood. Motivated by the desire to avoid the inefficiencies of MCMC algorithms rejecting a considerable number of expensive proposals, this thesis has developed an alternative inference framework based on adaptive multiple importance sampling (AMIS). In particular, this thesis studies the application of

AMIS for Gaussian processes in the case of a Gaussian likelihood, and proposes a novel pseudo-marginal-based AMIS (PM-AMIS) algorithm for non-Gaussian likelihoods, where the marginal likelihood is unbiasedly estimated. Experiments on benchmark data sets show that the proposed framework outperforms the MCMC-based inference of GP covariance parameters in a wide range of scenarios.

The PM-AMIS classifier - based on Gaussian processes with a newly designed group-automatic relevance determination (G-ARD) kernel - has been applied to predict whether a Flickr user is perceived to be above the median or not with respect to each of the Big-Five personality traits. The results show that, apart from the high prediction accuracies achieved (up to 79% depending on the trait), the parameters of the G-ARD kernel allow the identification of the groups of features that better account for the classification outcome and provide indications about cultural effects through their weight differences. Therefore, this demonstrates the value of the proposed non-parametric probabilistic framework for social signal processing.

Feature extraction in signal processing is dominated by various methods based on short time Fourier transform (STFT). Recently, Hilbert spectral analysis (HSA), a new representation of signal which is fundamentally different from STFT has been proposed. This thesis is also the first attempt to investigate the extraction of features from this newly proposed HSA and its application in social signal processing. The experimental results reveal that, using features extracted from the Hilbert spectrum of voice data of female speakers, the prediction accuracy can be achieved by up to 81% when predicting their Big-Five personality traits, and hence show that HSA can work as an effective alternative to STFT for feature extraction in social signal processing.

ACKNOWLEDGEMENTS

The completion of this thesis would not have been possible without the support of many people.

First and foremost, I would like to express my deep gratitude to my supervisors, Dr Maurizio Filippone and Prof Alessandro Vinciarelli for their help, encouragement, friendship and all their great inputs and insights along the way.

I am also grateful to all my friends and colleagues in Glasgow who made my time more enjoyable and memorable.

Finally, I would like to thank my family (Ning in particular) for their love and support.

AUTHOR'S DECLARATION

I hereby declare that all of the work presented in this thesis was performed personally unless otherwise stated. No part of this work has been submitted for consideration as part of any other degree or award.

Xiaoyu Xiong

Table of Contents

List of Tables

List of Figures

Acronyms

AIS adaptive importance sampling.

AM-FM amplitude modulation - frequency modulation.

AMIS adaptive multiple importance sampling.

ARD automatic relevance determination.

EP expectation propagation.

GP Gaussian process.

GPs Gaussian processes.

HC harmonic correspondence.

HMC hybrid Monte Carlo.

HSA Hilbert spectral analysis.

HSA-IMF Hilbert Spectral analysis - intrinsic mode function.

HT Hilbert transform.

IA instantaneous amplitude.

IF instantaneous frequency.

IMF intrinsic mode function.

LA Laplace approximation.

MAMIS modified version of AMIS.

MCMC Markov chain Monte Carlo.

MH Metropolis-Hastings.

NUTS No-U-Turn sampler.

NUTSDA NUTS with dual averaging.

PM-AMIS pseudo-marginal AMIS.

PM-MH pseudo-marginal MH.

PMC population Monte Carlo.

RBF radial basis function.

SHC simple harmonic component.

SIR Sampling Importance Resampling.

SMC sequential Monte Carlo.

SS slice sampling.

SSP social signal processing.

STFT short time Fourier transform.

SVM support vector machine.

Chapter 1

Introduction

Our everyday life revolves around interactions with others. We communicate with people in most of our daily activities - both at work and at home. Therefore, we as human beings, are actually social animals [3]. Not only do social interactions constellate our daily life, they also play an important role in a variety of social media. For example, we watch political debates or talk-shows on TV, we like pictures that are posted on Flickr or Instagram and we write our comments on Facebook or Twitter. Thus social intelligence - the skill of managing social signals (e.g. turn taking and mirroring) and social behaviours (e.g. agreement, empathy and politeness) involved in a social interaction with others - is indispensable and affects significantly the success of our life [56]. Consequently, a large number of computing efforts have been made to develop automatic approaches to analyse social interactions and have made social signal processing (SSP) an emerging research and technological domain in the computing community [158, 159].

Social signal processing has attracted the interest of a wide range of scientific communities in, for example, psychology, computer vision and signal processing. It aims to automatically understand and interpret social signals generated during human-human and human-machine interactions [112, 158]. Automatic interpretation of social signals involves two fundamentally important aspects: feature extraction and machine learning.

So far, machine learning approaches applied to social signal processing have mainly focused on the mapping between detectable behavioural cues (e.g. facial expressions and prosody)

and social and psychological phenomena using parametric approaches (e.g. linear regression) or non-parametric models such as support vector machines (SVM). Section 6.2 presents an extensive survey conducted on the computational approaches employed in personality inference from social media data - an important domain of SSP. However, such parametric approaches fall short of taking into account any uncertainty as a result of model misspecification (e.g. imposing a linear model on the data even though it is not able to well explain the data) or the large number of behavioural cues compared to the available annotated data (over-fitting). In addition, despite its non-parametric formulation, being a non-probabilistic approach makes SVM deficient in quantifying uncertainty in predictions and it is also statistically not able to assess the predictive effect of different features of the inputs on predicting the outputs [41].

Nevertheless, an accurate quantification of uncertainty in predictions and being able to gauge the degree to which each feature of the inputs is relevant to the prediction outcome, are of great importance for social signal processing applications. For example, when a system of analysing facial expressions is used in medicine to monitor the pain level or anxiety of a patient or used in security to assert the credibility of a person [153], a model capable of accurately quantifying uncertainty would be very useful. In the case of automatically detecting the conflict level in political debates [75], it is desirable to get a direct interpretation of the relative influence of different behavioural cues on predictions so that changing the associated behaviours in future debates could change people's perception of conflict.

One way to tackle these limitations is through a Bayesian treatment of the problem and a relaxation of the parametric assumptions to allow for an unknown functional form [121] for the mapping.

Gaussian processes (GPs), a non-parametric Bayesian framework, have proved to be a successful class of statistical inference methods for data analysis in several applied domains, such as pattern recognition [10, 40, 121], neuro-imaging [41], signal processing [75], Bayesian optimisation [72], and emulation and calibration of computer codes [74].

GPs are attractive because they are flexible and highly descriptive models with a superior capability of quantifying uncertainty in predictions: their non-parametric formulation yields the possibility of flexibly modelling data and, formulated in probabilistic terms, their

Bayesian treatment allows the incorporation of confidence levels when making predictions. In addition, with a suitable parametrisation of the covariance function, they offer the possibility of gaining some insights into the application under study without explicit knowledge of the mapping between inputs and outputs. The application of GP models to SSP data [75] suggests that the flexibility and interpretability offered by such non-parametric models can greatly enrich our understanding and interpretation of human behaviours. These properties of GP models hinge on the parametrisation of the GP covariance function and on the way GP covariance parameters are optimised or inferred.

Therefore, it is necessary to accurately characterise the posterior distribution over covariance parameters and extend this source of uncertainty forward to predictions [40, 41, 99, 140]. This task, which is the focus of this thesis, is particularly challenging when dealing with GPs. Inference of GP covariance parameters in closed form is generally analytically intractable and, when resorting to standard inference methods, a complication arises from the difficulties associated with having to repeatedly compute the marginal likelihood (and possibly the gradient of its logarithm). The marginal likelihood is computable in the case of a Gaussian likelihood, but extremely costly because of the need to carry out a number of operations that is cubic with the number of input vectors. On the other hand, when the likelihood function is not Gaussian, e.g. in classification, in ordinal regression, or in Cox-processes, the marginal likelihood is not even computable analytically.

In response to these challenges, a large body of the literature has developed approximate inference methods [63, 80, 104, 108, 121, 164] which, although successful in many cases, give no guarantees on the amount of bias they introduce that may affect their ability to quantify uncertainty. With regards to quantifying uncertainty without introducing any bias, there have been attempts to employ Markov chain Monte Carlo (MCMC) techniques. We can broadly divide such attempts in works that propose parametrisation techniques [42, 96, 99, 154] or methods that carry out inference based on unbiased computations of the marginal likelihood [39, 40, 97]. Although these approaches proved successful in a variety of scenarios, employing MCMC algorithms may lead to inefficiencies; for instance, optimal acceptance rates for popular MCMC algorithms such as the Metropolis-Hastings (MH) algorithm (approximately 25% [125]) and the hybrid Monte Carlo (HMC) algorithm (approximately 65% [9, 101]) indicate that several expensive computations are wasted. Introducing adap-

tivity into MCMC proposal mechanisms to improve efficiency may lead to convergence issues [1].

Consequently, it is difficult to use GPs in real-world applications. In the case of GPs with a Gaussian likelihood, the marginal likelihood is computable but expensive (scaling with the cube of the number of data); is it possible to accurately quantify uncertainty (in terms of being able to sample from the posterior and make good predictions) while mitigating the effect of the inefficiencies of MCMC methods? In the case of non-Gaussian likelihoods, the marginal likelihood is not even computable; is it possible to do so when the marginal likelihood is not available analytically? Indeed, is it possible to effectively employ the non-parametric GP framework to analyse the SSP data? The work in this thesis mainly aims to offer solutions to these problems.

1.1 Aims

The aims of the thesis are as follows:

1. To develop a general framework to carry out Bayesian inference for GPs aimed at overcoming the limitations of MCMC methods, where expectations under the posterior distribution over covariance parameters are carried out by means of the adaptive multiple importance sampling (AMIS) algorithm [22]. The application of this framework to the Gaussian likelihood case, although novel, is relatively straightforward given that the likelihood is computable. This thesis extensively compared the sampling efficiency (in terms of convergence speed against computational complexity) of MCMC versus AMIS for GP regression where the marginal likelihood is computable and the experimental results showed that AMIS can achieve faster convergence speed in this case.

   In the case of non-Gaussian likelihoods, the inability to compute the likelihood exactly leads to proposing a novel version of AMIS where the likelihood is unbiasedly estimated. Inspired by the Pseudo-Marginal MCMC approaches [2], the pseudo-marginal AMIS (PM-AMIS) algorithm was proposed, and a theoretical analysis was provided showing under which conditions PM-AMIS yields expectations under the posterior

over GP covariance parameters without introducing any bias. The proposed PM-AMIS is an instance of the Importance Sampling squared (IS^2) algorithms [114, 145] that are gaining popularity as practical Bayesian inference methods. There was also a extensive comparison of the sampling efficiency of PM-AMIS versus PM-MCMC for GP classification where the marginal likelihood cannot be computed analytically. In this case, the experimental results showed that the convergence speed of PM-AMIS is also faster than that of PM-MCMC.

2. To explore whether the above proposed Bayesian GP framework is suitable for the analysis (description and interpretation) of scenarios in social signal processing (SSP). This is at the border between computing and psychology and it can be addressed with experiments on data collected during social interactions. In this thesis, the proposed PM-AMIS has been applied with a newly designed kernel - the *Group Automatic Relevance Determination* (G-ARD) - to classify personality traits of people from the online Flickr pictures. The results demonstrated the value of this proposal for SSP: this new methodology not only can accurately predict the personality traits of the Flickr users, it also has the major advantage of being able to identify visual characteristics of Flickr images that mostly influence the personality impression.

3. To examine alternative feature extraction method based on the Hilbert spectrum for social signal processing. Feature extraction in signal processing is dominated by various methods based on short time Fourier transform (STFT). Recently, Hilbert spectral analysis (HSA), a new representation of signal fundamentally different from STFT, has been proposed by [131] for signal processing. It is, therefore, proposed to investigate feature extraction from HSA and its application in social signal processing. In this thesis, the extraction of features from the Hilbert spectrum (HS) and STFT of voice data - fillers (sounds filling a pause in a conversation) of female speakers - was explored. The resulting features are called HS features and STFT features, respectively. Both the HS and STFT features were used to predict the Big-Five personality traits [132] of the female speakers. The results showed that the prediction accuracies achieved using HS features were competitive with those obtained using STFT features, and suggested an alternative feature extraction method for social signal processing.

1.2 Thesis Statement

GPs have gained popularity in data analysis because of their attractive properties - being non-parametric enables them to flexibly model data and being probabilistic makes them capable of quantifying uncertainty. In addition, different from SVM that is widely used in SSP, the parametrisation of the GP covariance function allows insights into the application under study. However, inferring GP covariance parameters is particularly challenging since the computation of their marginal likelihoods is very costly. Traditional inference techniques for GPs, such as MCMC, have inefficiency problems caused by their rejection of expensive proposals and potential over-estimation of the marginal likelihood. Although parametric approaches and SVM have been enormously used in SSP, their inherent limitations make them less able to model the data effectively. With HSA, despite being a new representation of signal, its practical application to SSP has never been explored. The core assertion of this thesis is that, using AMIS for inference of GP covariance parameters can mitigate the inefficiencies of the MCMC algorithms; GPs with the novel G-ARD kernel offers an efficient probabilistic framework for SSP which can improve our understanding and interpretation of human behaviours. HSA can work as a valid alternative to STFT for feature extraction in social signal processing.

This assertion is supported through a number of experiments. An extensive comparison between AMIS and MCMC in terms of convergence speed against computational complexity suggested that AMIS is competitive with MCMC algorithms when calculating expectations under the posterior distribution over GP covariance parameters. The results showed that AMIS is a valid alternative to MCMC algorithms even in the case of moderately large dimensional parameter spaces, which is common when employing richly parametrised covariances (e.g. automatic relevance determination (ARD) covariances [90]). Given that importance sampling-based inference methods, unlike MCMC algorithms, are inherently parallel, experimental results in the thesis suggested a promising direction to accelerate the inference of GP covariance parameters.

When GPs with AMIS are applied to conduct personality inference from the pictures that users have tagged as favourite on Flickr, aside from the high prediction accuracies achieved,

the novel G-ARD kernel allows the idenfification of visual characteristics that better account for the prediction outcome while detecting cultural differences between the UK and Asian personality assessors.

Using features extracted from the Hilbert spectrum of fillers of female speakers, the prediction accuracies of personality traits of female speakers are comparable with those achieved using features extracted from conventional STFT output. The results suggested that HSA is an effective alternative feature extraction approach for SSP.

1.3 List of Contributing Papers

The work described in this thesis has led to one journal paper and one conference paper as follows:

- X. Xiong, V. Šmídl, and M. Filippone. Adaptive multiple importance sampling for Gaussian processes. *Journal of Statistical Computation and Simulation*, 87(8):1644–1665, 2017
- X. Xiong, M. Filippone, and A. Vinciarelli. Looking good with Flickr Faves: Gaussian processes for finding difference makers in personality impressions. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 412–415, 2016

Chapters 1 to 5 of this thesis are based on the first paper. Chapter 6 contains elements of the second paper. The analyses provided in this thesis are expanded treatments of the work in the papers, together with some additional unpublished research (Chapter 7).

1.4 Main Contributions

The main contributions of this thesis are as follows:

- The application of AMIS to infer GP covariance parameters with any likelihood (Chapter 4).

- A theoretical analysis of PM-AMIS (Chapter 4).
- An extensive comparison of convergence speed with respect to computational complexity of AMIS versus MCMC methods (Chapter 5).
- An application of PM-AMIS with the novel G-ARD kernel to perform personality inference from online pictures that has demonstrated the value of the proposed non-parametric probabilistic framework for SSP (Chapter 6).
- The first experimental exploration of feature extraction from the Hilbert spectrum that provides an alternative feature extraction method for SSP (Chapter 7).

Table 1.1 illustrates where the work fits in the literature of Bayesian inference for GP covariance parameters and beyond.

Table 1.1 Schematic representation of where the proposed contribution (highlighted in bold red) fits within the literature.

Inference	Models	(Marginal) Likelihood		Reparameterizations
		Computable	Estimated	
MCMC	Others	[98]	[2]	[111]
	GPs	[99]	[40]	[42]
AMIS	Others	[22]	**PM-AMIS**	–
	GPs	**AMIS for GPs**		–

Notes: This thesis proposes AMIS for Gaussian processes and PM-AMIS and studies its application to Gaussian processes; the latter can be employed whenever an unbiased estimate of the (marginal) likelihood is available. The list of references is not exhaustive but illustrates some of the key works and reviews in this field.

1.5 Thesis Walkthrough

The rest of the thesis is organised into 7 chapters.

Chapter 2, *Bayesian Gaussian Processes*, provides an overview of GP in Section 2.1, with Gaussian likelihood and non-Gaussian likelihood described in Sections 2.1.1 and 2.1.2, respectively. In Section 2.1.2, two popular approximation approaches - Laplace approximation and expectation propagation - are also described. Section 2.2 presents Bayesian inference

of GP covariance parameters. In Section 2.3, predictions under Gaussian likelihood and non-Gaussian likelihood are examined. Section 2.4 concludes Chapter 2.

Chapter3, *MCMC methods*, reviews the state-of-the-art MCMC methods: slice sampling (Section 3.1), hybrid Monte Carlo including the No-U-Turn Sampler (Section 3.2), Metropolis-Hastings (Section 3.3) and pseudo-marginal MCMC (Section 3.4). Section 3.5 concludes Chapter 3.

Chapter 4, *Adaptive Monte Carlo*, initially reviews adaptive MCMC in Section 4.1, then presents AMIS for GPs in Section 4.2. Section 4.3 describes the proposed PM-AMIS. Section 4.4 concludes Chapter 4.

Chapter 5, *Experiments and Results*, reports on the experiments and results of convergence analysis of AMIS versus MCMC for both the GP regression and classification cases.

Chapter 6, *Gaussian Processes for Finding Difference Makers in Personality Impressions - an Application of PM-AMIS*, examines the application of PM-AMIS to personality analysis, an important area of SSP. Section 6.1 introduces human behaviour analysis from social media data. Section 6.2 provides an extensive survey conducted on the computational approaches adopted in personality analysis from social media data. In section 6.3, data used in the experiments, personality and its assessment are described. Section 6.4 details the feature extraction from the Flickr pictures. In section 6.5, two classification approaches - PM-AMIS with the proposed novel G-ARD kernel and SVM, and the experimental setup are described. Section 6.6 reports on the experiments and results. Section 6.7 concludes Chapter 6.

Chapter 7, *Feature Extraction Using Hilbert Spectral Analysis*, investigates the application of HSA to SSP, examining the predictive effect of features extracted from the Hilbert spectrum of fillers of female speakers. Section 7.1 describes the background of HSA and Section 7.2 presents the HSA algorithm used in the experiments. Section 7.3 reports on the experiments and results of feature extraction from the Hilbert spectrum. Section 7.4 concludes Chapter 7.

Chapter 8, *Conclusions*, gives a summary of the results and provides suggestions for future work.

Chapter 2

Bayesian Gaussian Processes

This chapter reviews Gaussian processes (GPs) and begins by examining the reason for the adoption of GP in supervised learning, its definition and two common GP covariance functions in Section 2.1. It describes the marginal likelihoods for the cases of Gaussian and non-Gaussian likelihoods in Section 2.1.1 and Section 2.1.2, respectively. In Section 2.1.2, in order to integrate out the latent variables introduced by the non-Gaussian likelihood, two popular approximation approaches - Laplace approximation (LA) and expectation propagation (EP) - are also described.

Section 2.2 presents the Bayesian inference of GP covariance parameters; in particular, two stochastic approaches (based on MCMC and importance sampling) to compute the expectations under the posterior over GP covariance parameters, which are the focuses of this thesis, are examined.

Section 2.3 derives the predictive distributions under the GP framework with the cases of the Gaussian and non-Gaussian likelihoods described in Section 2.3.1 and Section 2.3.2 respectively. In Section 2.3.1, an example of GP regression is also presented, showing GP's capability of quantification of uncertainty. In Section 2.3.2, where the likelihood is non-Gaussian, predictive distribution under deterministic approximations (LA and EP) as well as a fully Bayesian treatment for prediction, are discussed.

Section 2.4 ends with a summary of this chapter.

2.1 Gaussian Processes

The core of supervised learning is to find the input-output mappings from the observed data (in particular training data). Given a training data set, the aim is to find an underlying function that makes predictions for all possible input values. Two common approaches have been employed to deal with the problem of supervised learning. The first is to restrict the class of functions, e.g. by only considering linear mappings between input and output. The drawback of this approach is its limited expressiveness: we may make wrong assumptions about the model (e.g. the data cannot be well explained by a linear model but we impose a linear model on the data) and hence the predictions based on it will be poor. One solution to this problem is to increase the flexibility of the class of functions, for example, by projecting the inputs into high dimensional spaces using a set of basis functions. However, there is, therefore the possible risk of overfitting and it is important to face the problem of how to choose the basis functions [121].

The other approach takes a Bayesian view and aims to assign a prior probability to a general class of functions, with higher probabilities given to functions considered to be more likely. A serious problem with this approach is its computational intractability: there is an infinite set of possible functions, so it is not possible to compute with this set in finite time. This is where Gaussian processes come into play. A Gaussian process (GP) is a generalisation of the Gaussian probability distribution. Compared to a probability distribution, defined over random variables which are scalars (for univariate distributions) or vectors (for multivariate distributions), a stochastic process controls the properties of functions. That is, a GP describes a distribution over functions, where a function can be seen as an infinite dimensional vector, of which each component specifies the function value $f(\mathbf{x})$ at a particular input $\mathbf{x}$. A GP framework solves this problem of computational intractability by the attractive consistency property of multivariate Gaussian distribution - the marginal and conditional densities of a multivariate Gaussian are also Gaussian, which enables us to focus exclusively on the variables of interest whilst ignoring the rest when making an inference. That is to say, under the GP framework, an inference achieved by just considering functions evaluated at a finite number of inputs will be identical to that derived by taking into account the infinitely many other points.

A formal definition of GP is given by [121] as follows:

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

Here random variables correspond to the function values $f(\mathbf{x})$ at all possbile locations of $\mathbf{x}$.

A GP is specified by its mean function and covariance function, which are defined by

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}(f(\mathbf{x})) \\ k(\mathbf{x}_i, \mathbf{x}_j) &= \mathbb{E}\left[(f(\mathbf{x}_i) - m(\mathbf{x}_i))(f(\mathbf{x}_j) - m(\mathbf{x}_j))\right] \end{aligned} \tag{2.1}$$

where the mean function $m(\mathbf{x})$ is usually taken to be zero for notational simplicity.

Note that the above definition of GP automatically implies the marginalisation property of eq. (C.4), which means that examination of a larger set of variables does not change the distribution of a smaller set (see Appendix C).

Here is an example of a supervised learning scenario. Let $\mathbf{X}$ be a set of n input vectors $\mathbf{x}_i \in \mathbb{R}^d (1 \leq i \leq n)$, and let $\mathbf{y}$ be the vector consisting of the corresponding labels y_i. In most GP models, the labels are assumed to be conditionally independent given a set of n latent variables. Such latent variables are modeled as realisations of a function $f(\mathbf{x})$ at the input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, i.e. $\mathbf{f} = \{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)\}$. Latent variables are used to express the likelihood function, which under the assumption of independence becomes $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} p(y_i \mid f_i)$, where $p(y_i \mid f_i)$ depends on the data being modelled (e.g. Gaussian for regression, Bernoulli for probit classification with probability $P(y_i = 1) = \Phi(f(\mathbf{x}_i))$ where Φ is defined as the cumulative normal distribution).

What characterises GP models is the way the latent variables are specified. In particular, it is assumed that the function $f(\mathbf{x})$ is distributed as a GP, which implies that the latent function values $\mathbf{f}$ are jointly distributed as a Gaussian $p(\mathbf{f} \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} \mid \mathbf{0}, K)$, where K is the covariance matrix. The entries of the covariance matrix K are specified by a covariance (kernel) function with hyperparameters $\boldsymbol{\theta}$. The Gaussian distribution $p(\mathbf{f} \mid \boldsymbol{\theta})$ is usually called a GP prior. A similarity matrix K is chosen so that, when the approximate functions are fitted to the data, it is possible to make sure that if two input values are close by, the

corresponding outputs are also close by.

In this thesis, two different covariance functions have been considered. The first is the radial basis function (RBF) defined as:

$$\text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp\left\{-\frac{1}{\tau^2} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2\right\} \tag{2.2}$$

The parameter τ defines the characteristic length-scale of the interaction between the input vectors, while σ represents the marginal variance for each latent variable. Note that the covariance between the outputs $(f(\mathbf{x}_i), f(\mathbf{x}_j))$ is written as a function of the inputs $(\mathbf{x}_i, \mathbf{x}_j)$. It can be shown (see Section 4.3.1 of [121]) that the RBF covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions.

The second is the ARD covariance, which takes the form:

$$\text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp\left\{-\sum_{r=1}^{d} \frac{1}{{\tau_r}^2}(\mathbf{x}_{i(r)} - \mathbf{x}_{j(r)})^2\right\} \tag{2.3}$$

The advantage of the ARD covariance is that it accounts for the influence of each feature on the mapping between inputs and labels, with smaller values of parameters $(\tau_1, ..., \tau_d)$ indicating a higher influence of the corresponding features [75]: when the length-scale is very large, the covariance will become nearly independent of that input, effectively eliminating it from the inference [121]. For simplicity of notation, in the remainder of the thesis the vector of all covariance parameters will be denoted by $\boldsymbol{\theta}$.

The length-scale parameters τ (RBF) and $\tau_1, ..., \tau_d$ (ARD) can be thought of as roughly the distance to be moved in input space before the function value can change significantly [121]. Figure 2.1 shows the RBF kernel $k(x, 0) = \exp(-\frac{x^2}{\tau^2})$ with length-scales ranging from 0.2 to 2.0 and three samples drawn from the GP priors with the corresponding length-scales. As can be seen from the figure, larger length-scales give smoother functions.

When making predictions, using a point estimate of $\boldsymbol{\theta}$ has been reported to potentially underestimate the uncertainty in predictions or yield inaccurate assessment of the relative influence of different features [10, 40, 41]. Therefore, a Bayesian approach is usually adopted to overcome these limitations, which entails characterising the posterior distribution over covariance

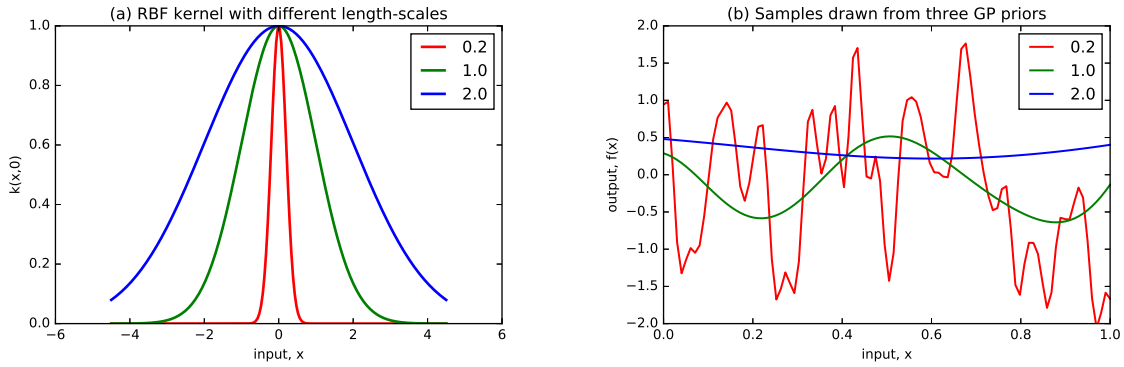


Figure 2.1 Panel (a) RBF kernel $k(x, 0) = \exp(-\frac{x^2}{\tau^2})$ with different length-scales τ. Panel (b) shows three samples drawn from the GP priors with the corresponding length-scales.

parameters. In order to do so, it is necessary to employ methods, such as MCMC, that require computing the marginal likelihood every time $\boldsymbol{\theta}$ is updated. It is now time to discuss the challenges associated with the computation of the marginal likelihood for the particular case of a Gaussian likelihood and the more general case of non-Gaussian likelihoods.

2.1.1 Gaussian Likelihood

In the GP regression setting, the observations $\mathbf{y}$ are modeled to be Gaussian distributed with a mean of $\mathbf{f}$ (latent variables) and covariance λI

$$p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \lambda I) \tag{2.4}$$

where I denotes the identity matrix, and λ is the variance of the Gaussian noise on the observations. In this setting, the likelihood $p(\mathbf{y} \mid \mathbf{f})$ and the GP prior $p(\mathbf{f} \mid \boldsymbol{\theta})$ form a conjugate pair, so latent variables can be integrated out of the model using eq. (C.10), leading to

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \boldsymbol{\theta}) d\mathbf{f} = \mathcal{N}(\mathbf{y} \mid \mathbf{0}, C) \tag{2.5}$$

where $C = K + \lambda I$. This yields the log-marginal likelihood

$$\log[p(\mathbf{y}|\boldsymbol{\theta})] = -\frac{1}{2}\log|C| - \frac{1}{2}\mathbf{y}^\top C^{-1}\mathbf{y} + \text{const.}$$

in closed form. Although computable, the log-marginal likelihood requires computing the log-determinant of C and solving a linear system involving C. These calculations are usually carried out by factorising the matrix $\mathbf{C}$ using the Cholesky decomposition, giving $C = LL^\top$, with $\mathbf{L}$ being the lower triangular. The Cholesky algorithm requires $O(n^3)$ operations, but subsequently computing the terms of the marginal likelihood requires at most $O(n^2)$ operations [121].

2.1.2 Non-Gaussian Likelihoods

In the case of non-Gaussian likelihoods, the likelihood $p(\mathbf{y} \mid \mathbf{f})$ and the GP prior $p(\mathbf{f} \mid \boldsymbol{\theta})$ are no longer conjugate. As a consequence, it is not possible to solve the integral needed to integrate out the latent variables

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f} \tag{2.6}$$

and this requires an approximation. A notable example is GP probit classification, which is explored in detail in this thesis. In this case, the observations $\mathbf{y}$ are assumed to be Bernoulli distributed with success probability given by [121]:

$$p(y_i \mid f_i) = \Phi(y_i f_i) \tag{2.7}$$

For GPs with non-Gaussian likelihoods, there have been several proposals on how to carry out approximation to integrate out the latent variables, or to avoid approximations altogether. The most popular approximations are the Laplace approximation (LA) [142] and expectation propagation (EP) [80, 121]. The following sections present a brief introduction of the LA and EP algorithms.

Laplace Approximation (LA)

The Laplace approximation approximates the target distribution of interest (the unnormalised posterior $p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \boldsymbol{\theta})$ in this case) by a Gaussian. The logarithm of the unnormalised

posterior is defined as:

$$\Psi(\mathbf{f}) = \log p(\mathbf{y} \mid \mathbf{f}) + \log p(\mathbf{f} \mid \boldsymbol{\theta}) \tag{2.8}$$

After applying a second order Taylor expansion to $\Psi(\mathbf{f})$, the following Gaussian approximation is obtained:

$$q(\mathbf{f} \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} \mid \hat{\mathbf{f}}, \hat{\Sigma}) \tag{2.9}$$

with the mean being the mode of $\Psi(\mathbf{f})$:

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \Psi(\mathbf{f}) \tag{2.10}$$

and the covariance determined by :

$$\hat{\Sigma} = -(\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \Psi(\hat{\mathbf{f}}))^{-1} \tag{2.11}$$

that is, minus the inverse Hessian of $\Psi(\mathbf{f})$ evaluated at the mode [142].

Solving the maximisation problem in eq. (2.10) involves an iterative procedure based on the following Newton-Raphson formula:

$$\mathbf{f}_{\mathbf{new}} = \mathbf{f} - (\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \Psi(\mathbf{f}))^{-1} \nabla_{\mathbf{f}} \Psi(\mathbf{f}) \tag{2.12}$$

Next is the derivation of the gradient ($\nabla_{\mathbf{f}} \Psi(\mathbf{f})$) and Hessian ($\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \Psi(\mathbf{f})$). Recall that under the GP assumption, the latent variables $\mathbf{f}$ are Gaussian distributed with the density:

$$p(\mathbf{f} \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} \mid \mathbf{0}, K) \tag{2.13}$$

where K is the covariance matrix with hyperparameters $\boldsymbol{\theta}$. The corresponding logarithm of $p(\mathbf{f} \mid \boldsymbol{\theta})$ takes the form:

$$\log[p(\mathbf{f}|\boldsymbol{\theta})] = -\frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi \tag{2.14}$$

where n is the number of data points.

Substituting eq. (2.14) into eq. (2.8) gives:

$$\Psi(\mathbf{f}) = \log p(\mathbf{y} \mid \mathbf{f}) - \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi \tag{2.15}$$

Differentiating eq. (2.15) w.r.t. $\mathbf{f}$ gives:

$$\nabla_{\mathbf{f}}\Psi(\mathbf{f}) = \nabla_{\mathbf{f}}\log p(\mathbf{y} \mid \mathbf{f}) - K^{-1}\mathbf{f} \tag{2.16}$$

$$\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\Psi(\mathbf{f}) = \nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\log p(\mathbf{y} \mid \mathbf{f}) - K^{-1} \tag{2.17}$$

It should be noted that if the likelihood $p(\mathbf{y} \mid \mathbf{f})$ is log concave, such as the probit likelihood defined in eq. (2.7), the Hessian $\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\Psi(\mathbf{f})$ will be negative definite. Consequently, $\Psi(\mathbf{f})$ is concave and has a unique maximum. In practice, when implementing the Newton-Raphson update in eq. (2.12), the *matrix inversion lemma*, also known as the Woodbury formula (eq. (C.13)), is employed to avoid inverting K directly (full details can be found in Section 3.4 of [121]). In this case, only one $O(n^3)$ operation is needed at each iteration for the $n \times n$ matrix factorisation.

The logarithm of the approximate marginal likelihood under Laplace approximation is given by:

$$\log q(\mathbf{y} \mid X, \boldsymbol{\theta}) = -\frac{1}{2}\hat{\mathbf{f}}^\top K^{-1}\hat{\mathbf{f}} + \log p(\mathbf{y} \mid \hat{\mathbf{f}}) - \frac{1}{2}\log \left| I + W^{\frac{1}{2}} K W^{\frac{1}{2}} \right| \tag{2.18}$$

where $W = -\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\log p(\mathbf{y} \mid \hat{\mathbf{f}})$.

One drawback of Laplace approximation is that the Hessian ($\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\Psi(\hat{\mathbf{f}})$) may give a poor approximation of the true shape of the target distribution (the posterior). Laplace approximation assumes the posterior has elliptical contours, while the peak of it could be skewed, or could be much narrower or broader than indicated by the Hessian [121].

Expectation Propagation (EP)

Expectation Propagation makes the assumption that each probit likelihood can be approximated by a local likelihood approximation, an unnormalised Gaussian function of f_i in the

form:

$$p(y_i \mid f_i) \simeq t_i(f_i \mid \tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma_i}^2) = \tilde{Z}_i \mathcal{N}(f_i \mid \tilde{\mu}_i, \tilde{\sigma_i}^2) \tag{2.19}$$

where $\tilde{Z}_i$, $\tilde{\mu}_i$, $\tilde{\sigma_i}^2$ are *site parameters*.

Approximating each individual term of the likelihood by a Gaussian implies the approximate likelihood is a multivariate Gaussian:

$$\mathcal{N}(\mathbf{f} \mid \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}) \prod_{i=1}^{n} \tilde{Z}_i \tag{2.20}$$

with $\tilde{\boldsymbol{\mu}}_i = \tilde{\mu}_i$ and $\tilde{\Sigma}_{ii} = \tilde{\sigma_i}^2$.

With this approximation, the posterior $p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$ is approximated by a Gaussian:

$$\begin{aligned} &q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma) \qquad \text{with} \\ &\boldsymbol{\mu} = \Sigma \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\mu}} \quad \text{and} \quad \Sigma = (K^{-1} + \tilde{\Sigma}^{-1})^{-1} \end{aligned} \tag{2.21}$$

Following the Gaussian identity of eq. (C.4), the marginal approximate postetior is given by

$$q(f_i \mid \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(f_i \mid \mu_i, \sigma_i^2) \tag{2.22}$$

where $\mu_i = \boldsymbol{\mu}_i$ and $\sigma_i^2 = \Sigma_{ii}$.

The main characteristic of the EP alogorithm is the way the site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$, $\tilde{\sigma_i}^2$ are optimised. Looping through the n factors approximating the likelihood, the EP algortihm optimises the three parameters of each factor t_i sequentially. Specifically, the optimisation involves the iteration of the following three steps. First, compute the approximate *cavity distribution* by leaving out the ith factor from $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$:

$$q_{-i}(f_i \mid \mathbf{y}, \boldsymbol{\theta}) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta}) \prod_{j \neq i} t(f_j) d\mathbf{f}_j \propto \mathcal{N}(f_i \mid \mu_{-i}, \sigma_{-i}^2) \tag{2.23}$$

Multiplying both sides of eq. (2.23) by $t(f_i)$ gives

$$q_{-i}(f_i \mid \mathbf{y}, \boldsymbol{\theta})t(f_i) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta}) \prod_{j \neq i} t(f_j)t(f_i)d\mathbf{f}_j \propto \mathcal{N}(f_i \mid \mu_i, \sigma_i^2) \tag{2.24}$$

Gaussian identity of eq. (C.6) gives the parameters:

$$\mu_{-i} = \sigma_{-i}^2(\sigma_i^{-2}\mu_i - \tilde{\sigma}_i^{-2}\tilde{\mu}_i) \quad \text{and} \quad \sigma_{-i}^2 = (\sigma_i^{-2} - \tilde{\sigma}_i^{-2})^{-1} \tag{2.25}$$

Secondly, find the Gaussian marginal (unnormalised) which closely approximates the product of the cavity distribution and the exact ith likelihood:

$$\hat{q}(f_i) = \hat{Z}_i \mathcal{N}(f_i \mid \hat{\mu}_i, \hat{\sigma}_i^2) \simeq q_{-i}(f_i)p(y_i \mid f_i) \tag{2.26}$$

The parameters of $\hat{q}(f_i)$ are found by minimising the Kullback-Leibler divergence :

$$\mathrm{KL}\Big(q_{-i}(f_i)p(y_i \mid f_i) \;\Big|\Big|\; \hat{q}(f_i)\Big) \tag{2.27}$$

which in practice is achieved by means of moments matching. In particular, $\hat{Z}_i$, $\hat{\mu}_i$, $\hat{\sigma}_i^2$ corresponds to the zero-th, first and second moments of $q_{-i}(f_i)p(y_i \mid f_i)$, respectively. The derivation of the moments can be found in Section 3.9 of [121]. The corresponding moments of the posterior marginal are

$$\begin{aligned} &\hat{Z}_i = \Phi(z_i) \qquad \hat{\mu}_i = \mu_{-i} + \frac{y_i \sigma_{-i}^2 \mathcal{N}(z_i)}{\Phi(z_i)\sqrt{1+\sigma_{-i}^2}} \\ &\hat{\sigma}_i^2 = \sigma_{-i}^2 - \frac{\sigma_{-i}^4 \mathcal{N}(z_i)}{(1+\sigma_{-i}^2)\Phi(z_i)}\Big(z_i + \frac{\mathcal{N}(z_i)}{\Phi(z_i)}\Big) \quad \text{with} \quad z_i = \frac{y_i \mu_{-i}}{\sqrt{1+\sigma_{-i}^2}} \end{aligned} \tag{2.28}$$

Thirdly, update the parameters of t_i as follows:

$$\begin{aligned} &\tilde{\mu}_i = \tilde{\sigma}_i^2(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_{-i}^{-2}\mu_{-i}) \qquad \tilde{\sigma}_i^2 = (\hat{\sigma}_i^{-2} - \sigma_{-i}^{-2})^{-1} \\ &\tilde{Z}_i = \hat{Z}_i \sqrt{2\pi}\sqrt{\sigma_{-i}^2 + \tilde{\sigma}_i^2} \exp\Big(\frac{1}{2}(\mu_{-i} - \tilde{\mu}_i)^2 \;/\; (\sigma_{-i}^2 + \tilde{\sigma}_i^2)\Big) \end{aligned} \tag{2.29}$$

Equation (2.29) can be verified by multiplying the cavity distribution $q_{-i}(f_i)$ by the local likelihood approximation t_i using eq. (C.6) to obtain eq. (2.28).

One iteration of the above three steps requires five operations in $O(n^3)$, which makes EP approximation computationally expensive.

The logarithm of the approximate marginal likelihood under EP approximation takes the form:

$$\log q(\mathbf{y} \mid X, \boldsymbol{\theta}) = -\frac{1}{2}\log|K + \tilde{\Sigma}| - \frac{1}{2}\tilde{\boldsymbol{\mu}}^{\mathbf{T}}(K + \tilde{\Sigma})^{-1}\tilde{\boldsymbol{\mu}} \\ + \sum_{i=1}^{n} \log \Phi(\frac{y_i \mu_{-i}}{\sqrt{1+\sigma_{-i}^2}}) + \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_{-i}^2 + \tilde{\sigma}_i^2) + \sum_{i=1}^{n}\frac{(\mu_{-i} - \tilde{\mu}_i)^2}{2(\sigma_{-i}^2 + \tilde{\sigma}_i^2)} \tag{2.30}$$

Although the EP algorithm is not guaranteed to converge in general, it has been reported that EP always converges for Gaussian process classification with probit likelihood [80], and no convergence issues have been reported in the literature [40]. Furthermore, despite the high computational cost involved, EP is usually the preferred method in terms of accuracy compared to other approximation approaches [104].

As this thesis focuses on stochastic methods (based on MCMC and importance sampling) to integrate out the latent variables and covariance parameters, the following section will consider these two stochastic approaches.

2.2 Bayesian Inference of Covariance Parameters

For simplicity of notation, the posterior distribution over covariance parameters is denoted by:

$$\pi(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2.31}$$

where $p(\boldsymbol{\theta})$ encodes any prior knowledge on the parameters $\boldsymbol{\theta}$. Within the Bayesian framework, there is usually interest in calculating expectations of functions of $\boldsymbol{\theta}$ with respect to the posterior distribution, i.e. $E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})]$. For instance, setting $h(\boldsymbol{\theta}) = p(y_\star \mid \boldsymbol{\theta}, \mathbf{x}_\star, \mathbf{y}, \mathbf{X})$ obtains the predictive distribution for the label $y_\star$ associated with a new input vector $\mathbf{x}_\star$.

The denominator needed to normalise the posterior distribution $\pi(\boldsymbol{\theta})$ is intractable, so it is

not possible to characterise the posterior distribution analytically. Despite this complication, it is possible to resort to a Monte Carlo approximation to compute expectations under the posterior distribution of $\boldsymbol{\theta}$:

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] \simeq \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)}) \tag{2.32}$$

where $\boldsymbol{\theta}^{(i)}$ denotes the ith of N samples from $\pi(\boldsymbol{\theta})$. However, as it is generally not feasible to draw samples from $\pi(\boldsymbol{\theta})$ directly, it is necessary to resort to MCMC methods to generate samples from the posterior $\pi(\boldsymbol{\theta})$.

An alternative way to compute expectations is by means of importance sampling, which takes the following form:

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{2.33}$$

where $q(\boldsymbol{\theta})$ is the importance distribution. The corresponding Monte Carlo approximation is of the form:

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] \simeq \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)})\frac{\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \tag{2.34}$$

where the samples $\boldsymbol{\theta}^{(i)}$ are now drawn from the importance sampling distribution $q(\boldsymbol{\theta})$. The key to making this Monte Carlo estimator accurate is to choose $q(\boldsymbol{\theta})$ to be similar to the function that needs to be integrated, that is $h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. It is easy to verify that when this is the case, the variance of the importance sampling estimator is zero. Therefore, the success of importance sampling relies on constructing a tractable importance distribution $q(\boldsymbol{\theta})$ that well approximates $h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. In the remainder of this thesis, methods that adaptively construct $q(\boldsymbol{\theta})$ are studied and evaluated.

Both Monte Carlo approximations in eq. (2.32) and eq. (2.34) converge to the desired expectation, and in practice, they can estimate the desired integral to a given level of precision [51, 44]. The experimental part of this thesis (Chapter 5) is devoted to the study of the convergence properties of the expectation $E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})]$ with respect to the computational effort needed to carry out the Monte Carlo approximations in eq. (2.32) and eq. (2.34).

As the general task in supervised learning is to compute the expectations with respect to the

predictive distribution for the output label $y_\star$ given a new input $\mathbf{x}_\star$, the following section will present the predictive distributions under the GP framework. In particular, predictive distributions with the Gaussian and non-Gaussian likelihoods will be discussed, respectively.

2.3 Predictions

2.3.1 Predictions Under Gaussian Likelihood

Given a new input $\mathbf{x}_\star$ with the corresponding label $y_\star$, following eq. (2.5) gives:

$$p(\mathbf{y}, y_\star \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}, y_\star \mid \mathbf{0}, C') \tag{2.35}$$

where C' can be partitioned into

$$C' = \begin{bmatrix} C & \mathbf{k}_\star \\ \mathbf{k}_\star{}^\top & c_{\star\star} \end{bmatrix} \tag{2.36}$$

with $c_{\star\star} = k(\mathbf{x}_\star, \mathbf{x}_\star) + \lambda$ and $\mathbf{k}_\star$ being the vector with elements $k(\mathbf{x}_i, \mathbf{x}_\star)$ for $i = 1, ..., n$.

Following the Gaussian identity of eq. (C.5) gives:

$$\begin{aligned} &p(y_\star \mid \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(y_\star \mid \mu_\star, \beta_\star^2) \qquad \text{with} \\ &\mu_\star = \mathbf{k}_\star{}^\top C^{-1}\mathbf{y} \quad \text{and} \quad \beta_\star^2 = c_{\star\star} - \mathbf{k}_\star{}^\top C^{-1}\mathbf{k}_\star \end{aligned} \tag{2.37}$$

Next is an example of GP regression where the likelihood is Gaussian.

Figure 2.2 shows samples drawn from a GP prior and posterior resulting from a GP regression. Panel (a) shows the zero mean function (represented by the solid blue line) of the prior and two functions (denoted by the red and green lines) drawn at random from the prior. Panel (b) shows the predictive mean function (represented by the solid blue line) of the posterior and two random functions (denoted by the red and green lines) drawn from that posterior, i.e. the prior conditioned on the ten noisy observations indicated by the black dots. In both plots the shaded area represents the pointwise mean plus and minus twice the standard deviation

at each input value (corresponding to the 95% confidence region). The effectiveness of the Bayesian treatment can be seen from the figure: the uncertainty is significantly reduced close to the observations while the error bars grow rapidly away from the data points.

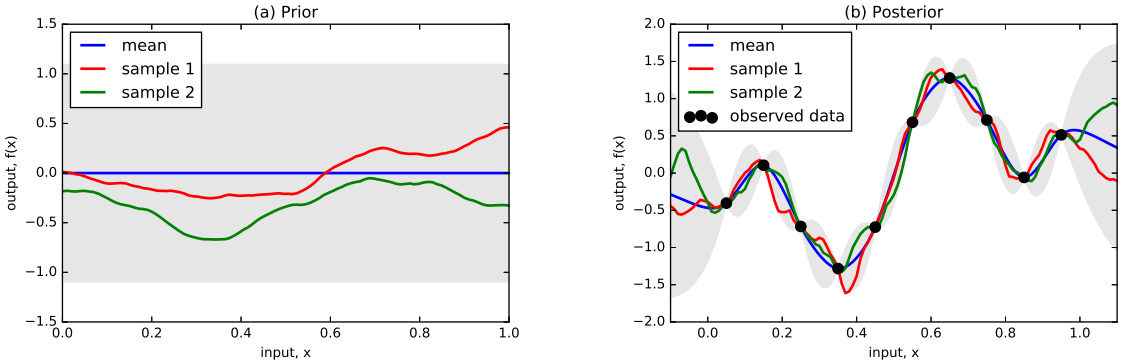


Figure 2.2 Panel (a) shows the zero mean function of the GP prior and two samples drawn at random from the prior. Panel (b) shows the situation after ten noisy datapoints have been observed. In both plots the shaded area denotes the pointwise mean plus and minus twice the standard deviation for each input value (corresponding to the 95% confidence region).

2.3.2 Predictions Under non-Gaussian Likelihood

As discussed earlier (Section 2.1.2), in the case of non-Gaussian likelihood, it is possible to resort to deterministic approximations such as LA and EP to obtain an approximate posterior of latent variables in order to exploit conjugacy. In the following section, how to derive the predictive distribution under such deterministic approximations will be examined.

Predictive distribution under deterministic approximations

The predictive distribution with respect to the approximate posterior $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$ in the case of Gaussian process classification is given by

$$p(y_\star \mid \mathbf{y}) = \int p(y_\star \mid f_\star) p(f_\star \mid \mathbf{f}, \boldsymbol{\theta}) q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta}) df_\star d\mathbf{f} \tag{2.38}$$

where $\boldsymbol{\theta}$ can be obtained by optimising the logarithm of the approximate marginal likelihood defined in eq. (2.30) and eq. (2.18) with respect to $\boldsymbol{\theta}$ or a sample from the approximate posterior up to a normalising constant $q(\mathbf{y} \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})$ using MCMC techniques.

Let K be the covariance matrix of the Gaussian prior evaluated at $\boldsymbol{\theta}$, $\mathbf{k}_\star$ be the vector the ith element of which is $k(\mathbf{x}_i, \mathbf{x}_\star \mid \boldsymbol{\theta})$ and $k_{\star\star} = k(\mathbf{x}_\star, \mathbf{x}_\star \mid \boldsymbol{\theta})$ where k is the kernel function defined in eq. (2.2) or eq. (2.3). According to GP definition, this gives

$$p(\mathbf{f}, f\star \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}, f\star \mid \mathbf{0}, K') \tag{2.39}$$

where K' can be partitioned as follows:

$$K' = \begin{bmatrix} K & \mathbf{k}_\star \\ \mathbf{k}_\star{}^\top & k_{\star\star} \end{bmatrix} \tag{2.40}$$

Then the Gaussian identity of eq. (C.5) gives:

$$\begin{aligned} &p(f_\star \mid \mathbf{f}, \boldsymbol{\theta}) = \mathcal{N}(f_\star \mid \mu_\star, \beta_\star^2) \qquad \text{with} \\ &\mu_\star = \mathbf{k}_\star{}^\top K^{-1}\mathbf{f} \quad \text{and} \quad \beta_\star^2 = k_{\star\star} - \mathbf{k}_\star{}^\top K^{-1}\mathbf{k}_\star \end{aligned} \tag{2.41}$$

When $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$ is approximated by a Gaussian $\mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)$ using the above LA or EP approximation (see Section 2.1.2), the integration with respect to $\mathbf{f}$ can be performed analytically (see Sections 3.4.2 and 3.6.1 of [121] for full details), yielding

$$\begin{aligned} &p(f_\star \mid \boldsymbol{\theta}) = \mathcal{N}(f_\star \mid m_\star, s_\star^2) \qquad \text{with} \\ &m_\star = \mathbf{k}_\star{}^\top K^{-1}\boldsymbol{\mu}_q \quad s_\star^2 = k_{\star\star} - \mathbf{k}_\star{}^\top K^{-1}\mathbf{k}_\star + \mathbf{k}_\star{}^\top K^{-1}\Sigma_q K^{-1}\mathbf{k}_\star \end{aligned} \tag{2.42}$$

Consequently the univariate integration with respect to $f_\star$ is computed as:

$$\int p(y_\star \mid f_\star)\mathcal{N}(f_\star \mid m_\star, s_\star^2)df_\star = \Phi\big(\frac{m_\star}{\sqrt{1+s_\star^2}}\big) \tag{2.43}$$

Compared to using deterministic approximations to integrate out latent variables when making predictions, a fully Bayesian treatment for prediction will now be described.

Fully Bayesian treatment for prediction

A fully Bayesian treatment aims to integrate out latent variables ($\mathbf{f}$) and hyperparameters ($\boldsymbol{\theta}$) of the covariance matrix of the GP prior:

$$p(y_\star \mid \mathbf{y}) = \int p(y_\star \mid f_\star) p(f_\star \mid \mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y}) df_\star d\mathbf{f} d\boldsymbol{\theta} \tag{2.44}$$

Define $h(\boldsymbol{\theta}, \mathbf{f}) = \int p(y_\star \mid f_\star) p(f_\star \mid \mathbf{f}, \boldsymbol{\theta}) df_\star$. As $f_\star$ is distributed as a Gaussian in the form of eq. (2.41), it can be integrated out analytically. Therefore, there are two ways to compute the expectation $p(y_\star \mid \mathbf{y})$ in eq. (2.44). One way is to use MCMC methods to sample from the posterior $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$ and compute the Monte Carlo estimate using eq. (2.32). The other is to use importance sampling to compute the Monte Carlo estimate using eq. (2.34).

However, it is not feasible to sample from the posterior $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$ by joint proposals because "*it is extremely unlikely to propose a set of latent variables and hyperparameters that are compatible with each other and observed data*" [40]. By employing

$$p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y}) = p(\mathbf{f} \mid \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \mid \mathbf{y}) \tag{2.45}$$

it is possible to solve this problem by first sampling $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta} \mid \mathbf{y})$ and then sampling $\mathbf{f}$ from $p(\mathbf{f} \mid \boldsymbol{\theta}, \mathbf{y})$ using MCMC or importance sampling.

2.4 Conclusion

In this chapter, there has been a description of the principle of Gaussian processes including two common covariance functions which are considered in this thesis. There has been a discussion of the computational challenges encountered when calculating the marginal likelihoods for the cases of both the Gaussian and non-Gaussian likelihoods. In the case of non-Gaussian likelihood, each iteration of the EP and LA algorithms needed to integrate out the latent variables requires five and one $O(n^3)$ operations respectively, further increasing the computational cost. The MCMC-based and importance sampling-based Monte Carlo approximations needed when conducting Bayesian inference of GP covariance parameters

have also been introduced. Finally, the predictive distributions in cases of Gaussian and non-Gaussian likelihoods have been derived. In the case of Gaussian likelihood, an example has been given showing the samples drawn from a GP prior and posterior respectively. In the case of non-Gaussian likelihood, the predictive distribution under deterministic approximations and a fully Bayesian treatment for prediction have been discussed.

This thesis focuses on stochastic approaches to compute the expectations under the posterior over GP covariance parameters, which are normally carried out by means of MCMC algorithms. The following will discuss the state-of-the-art MCMC approaches. However, as will be seen in Chapter 3, MCMC algorithms require repeated expensive computations of the marginal likelihood and the rejection of proposals leads to a waste of computations. In order to avoid the inefficiencies of MCMC algorithms, in Chapter 4, there will be a brief review of the adaptive MCMC approaches, and an exploration of the idea of using adaptive multiple importance sampling algorithms (AMIS) [22] to infer GP covariance parameters for the particular case of Gaussian likelihood; there will also be an examination of the pseudo-marginal AMIS proposed for the more general case of non-Gaussian likelihood. In chapter 5, there will be a report on the experiments and results of convergence analysis of AMIS versus MCMC for both the GP regression and classification cases. Chapter 6 presents an application of the pseudo-marginal AMIS for a personality classification problem in social signal processing.

Chapter 3

MCMC Methods

As described in Chapter 2, Markov chain Monte Carlo (MCMC) methods are normally employed to solve the intractability problem when inferring GP covariance parameters. This chapter will see a review of the state-of-the-art MCMC algorithms. Section 3.1 examines the slice sampling algorithms [100]. Section 3.2 discusses the hybrid Monte Carlo (HMC) approach [33, 98], with descriptions of its variants (No-U-Turn Sampler and No-U-Turn Sampler with Dual Averaging [66]) in Section 3.2.1 and Section 3.2.2 respectively. Section 3.3 describes the Metropolis-Hastings algorithms [62, 95]. The foregoing MCMC algorithms to draw samples from the posterior over GP covariance parameters require the marginal likelihood to be computed exactly, that is, the likelihood is Gaussian. In order to deal with cases where the marginal likelihood is non-Gaussian, it is possible to resort to the pseudo-marginal MCMC approach (in particular the pseudo-marginal MH in this thesis), which is described in Section 3.4. Section 3.5 concludes this chapter.

3.1 Slice Sampling - SS

Slice sampling [100], is a method in which the joint probability density of the auxiliary variable u and parameters of interest $\boldsymbol{\theta}$ takes the form $\pi(\boldsymbol{\theta}, u)$, such that

$$\pi(\boldsymbol{\theta}, u) = \begin{cases} 1/C & \text{if } 0 < u < f(\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $C = \int f(\boldsymbol{\theta})d\boldsymbol{\theta}$, and $f(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$, i.e the posterior up to a constant defined in eq. (2.31) (this definition of $f(\boldsymbol{\theta})$ will apply in the remainder of this chapter). Then the marginal distribution for $\boldsymbol{\theta}$ is given by:

$$\pi(\boldsymbol{\theta}) = \int_0^{f(\boldsymbol{\theta})} (1/C)du = f(\boldsymbol{\theta})/C \tag{3.2}$$

as desired. To sample from the target distribution $\pi(\boldsymbol{\theta})$, it is possible to sample from $\pi(\boldsymbol{\theta}, u)$ and then ignore u. Specifically, this is done by alternating uniform sampling of u from the vertical interval from 0 to the density $f(\boldsymbol{\theta})$ evaluated at the current state, with uniform sampling of $\boldsymbol{\theta}$ from the union of intervals which are the horizontal "slices" defined by the vertical position. Single-variable slice sampling is easily implemented and one can do component-wise univariate slice sampling for a multivariate distribution by updating each variable in turn. This approach is reported in [100] to be easier to implement than Gibbs sampling and be able to sample more efficiently than simple Metropolis scheme, because of its ability to adaptively choose the scale of changes appropriate to the region of the target distribution being sampled. For example, if the rough guess at the initial interval (an estimate for the scale of the horizontal slice) is too small compared to the true width of the target density, it can be expanded by "stepping out" or "doubling", whereas if the initial interval is too large, it can be shrunk by an efficient shrinkage procedure. More elaborate multivariate slice samplers can not only adapt to the scale of variables, but also to the dependencies between variables. Sampling efficiency can also be improved by the "over-relaxed" univariate slice sampling and the "reflective" multivariate slice sampling that can suppress random walks. Algorithm 1 gives the univariate slice sampling algorithms.

Algorithm 1 Univariate SS, adapted from [100].

Given $\boldsymbol{\theta} =$ the current point, $f(\boldsymbol{\theta})$, $w =$ estimate of the typical size of the step (horizontal slice) for creating interval

- The "stepping out" procedure for finding an interval around the current point:
  1. Determine the random vertical position $v = u_1 f(\boldsymbol{\theta})$ with $u_1 \sim U_{[0,1]}$
  2. Randomly place the interval around the current point

$$[\boldsymbol{\theta}_{min}, \boldsymbol{\theta}_{max}] \leftarrow [\boldsymbol{\theta} - u_2, \boldsymbol{\theta} + (w - u_2)] \qquad \text{with} \qquad u_2 \sim U_{[0,w]}$$

3. Expand the interval by "stepping out" until its ends are outside the slice:
   while $f(\boldsymbol{\theta}_{min}) > v$: $\boldsymbol{\theta}_{min} \leftarrow \boldsymbol{\theta}_{min} - w$
   while $f(\boldsymbol{\theta}_{max}) > v$: $\boldsymbol{\theta}_{max} \leftarrow \boldsymbol{\theta}_{max} + w$

- The "shrinkage" procedure for sampling from the interval:

4. **Repeat:**
   Sample a new point $\boldsymbol{\theta}' \sim U_{[\boldsymbol{\theta}_{min}, \boldsymbol{\theta}_{max}]}$
   if $f(\boldsymbol{\theta}') > v$, accept $\boldsymbol{\theta}'$ **then** exit loop
   else shrink the interval $[\boldsymbol{\theta}_{min}, \boldsymbol{\theta}_{max}]$
   if $\boldsymbol{\theta}' < \boldsymbol{\theta}$, $\boldsymbol{\theta}_{min} \leftarrow \boldsymbol{\theta}'$ **else** $\boldsymbol{\theta}_{max} \leftarrow \boldsymbol{\theta}'$

3.2 Hybrid Monte Carlo - HMC

HMC [33, 98] originated from Physics, where the Hamiltonian dynamics function is defined by the sum of a potential energy function of the position vector and a kinetic energy function of the momentum vector. When HMC is applied to obtaining samples from a target distribution, the parameters of interest $\boldsymbol{\theta}$, take the role of the position, and an auxiliary "momentum" variable $\mathbf{p}$, which is commonly assumed to be independently drawn from $\mathcal{N}(\mathbf{p} \mid 0, \mathbf{M})$, needs to be introduced. Thus the extended target distribution $\pi(\boldsymbol{\theta}, \mathbf{p})$ up to a constant takes the form $\exp(\mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p})$ where $\mathcal{L}(\boldsymbol{\theta})$ is the logarithm of $f(\boldsymbol{\theta})$, i.e., $\mathcal{L}(\boldsymbol{\theta}) = \log[f(\boldsymbol{\theta})]$. Consequently, the minus logarithm of the augmented target distribution plus some constant will give an analogy with the Hamiltonian:

$$\mathcal{H}(\boldsymbol{\theta}, \mathbf{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + \text{const.} \tag{3.3}$$

where $-\mathcal{L}(\boldsymbol{\theta})$ is the potential energy and $\frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}$ is the kinetic energy.

Generating a new sample of parameters by HMC involves three steps, as reported in Algorithm 2. The first is to draw $\mathbf{p}$ randomly from $\mathcal{N}(\mathbf{p} \mid 0, \mathbf{M})$, then to propose a new $\boldsymbol{\theta}_{(L)}, \mathbf{p}_{(L)}$ through a number of L reversible *leapfrog* steps (with step-size ϵ) that follow the Hamiltonian dynamics scheme. The *leapfrog* integrator (described by the Leapfrog function in Algorithm 2) gives the discrete-time simulation of the Hamiltonian dynamics scheme. As a result of this approximation by discretion, the energy is no longer conserved. Therefore, to ensure HMC samples from the correct invariant distribution, a Metropolis accept/reject step to accept the proposed $\boldsymbol{\theta}_{(L)}, \mathbf{p}_{(L)}$ with probability $\min\left\{1, \exp(-\mathcal{H}(\boldsymbol{\theta}_{(L)}, \mathbf{p}_{(L)}) + \mathcal{H}(\boldsymbol{\theta}, \mathbf{p}))\right\}$ is needed in the last step of HMC. The step-size ϵ and the number of steps L are often chosen randomly to ensure ergodicity, and exploiting the gradient gives HMC one major advantage of avoiding the random walk behaviour that occurs in MH.

Because of the "shear" transformations in the Leapfrog function, where the update of one variable $\boldsymbol{\theta}$ depends only on the other unchangeable variable $\mathbf{p}$ [101], the *leapfrog* integrator is volume-preserving. The crucial property of reversibility and preservation of volume of HMC makes the Metropolis proposal valid and hence ensures sampling from the invariant target distribution of interest. Another benefit of HMC is its better scalability with dimensionality compared to simple Metropolis approaches, details of which can be found in [101]. The choice of the number of steps L and step-size ϵ can heavily affect the performance of HMC; thus careful tuning of these two parameters is usually needed when applying HMC. Using knowledge of scales and correlation of the position variables can also improve the performance of HMC [101]. Specifically, this is achieved by transforming the position variables $\boldsymbol{\theta}$ to $\mathbf{L}^{-1}\boldsymbol{\theta}$ or using a mass matrix $\mathbf{M} = \boldsymbol{\Sigma}^{-1}$ or $(\mathrm{diag}(\boldsymbol{\Sigma}))^{-1}$, where $\boldsymbol{\Sigma}$ denotes the estimate of the covariance of $\boldsymbol{\theta}$, and $\mathbf{L}$ denotes the lower triangular of the Cholesky decomposition of $\boldsymbol{\Sigma}$. There will now be a discussion of NUTS which automatically tunes ϵ, L.

Algorithm 2 HMC.

Given $\boldsymbol{\theta}$ = the current point, $\mathcal{L}(\boldsymbol{\theta})$, ϵ = step-size, M = mass matrix, L = sample$[1, ..., L_{max}]$, $\nabla_{\boldsymbol{\theta}}$ = gradient with respect to $\boldsymbol{\theta}$

1. Sample $\mathbf{p} \sim \mathcal{N}(0, \mathbf{M})$

2. Set $\boldsymbol{\theta}'$, $\mathbf{p}' \leftarrow \text{Leapfrog}(\boldsymbol{\theta}, \mathbf{p}, \epsilon)$

   for $i = 2$ to L **do**

   Set $\boldsymbol{\theta}'$, $\mathbf{p}' \leftarrow \text{Leapfrog}(\boldsymbol{\theta}', \mathbf{p}', \epsilon)$

   end for

3. Draw $u \sim U_{[0,1]}$

   if $u < \min\left\{1, \exp(-\mathcal{H}(\boldsymbol{\theta}_{(L)}, \mathbf{p}_{(L)}) + \mathcal{H}(\boldsymbol{\theta}, \mathbf{p}))\right\}$, **return** $\boldsymbol{\theta}'$

   else return $\boldsymbol{\theta}$

function $\text{Leapfrog}(\boldsymbol{\theta}, \mathbf{p}, \epsilon)$

Set $\mathbf{p}' = \mathbf{p} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$

Set $\boldsymbol{\theta}' = \boldsymbol{\theta} + \epsilon\mathbf{M}^{-1}\mathbf{p}'$

Set $\mathbf{p}' = \mathbf{p}' + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}')$

return $\boldsymbol{\theta}', \mathbf{p}'$

3.2.1 No-U-Turn Sampler - NUTS

As mentioned above, the performance of HMC can be significantly influenced by the choice of ϵ and L. Too large an ϵ will lead to a very low acceptance rate, whereas too small an ϵ will result in waste of computation time and also the risk of undesirable random walk when L is not large. Choosing L can be problematic as well. When L is too large, by taking too many steps or looping back to where the position variable was before, it will waste a lot of expensive computations. When L is too small, the resulting random walk behaviour will cause slow exploration of the state and thus poor-mixing of the samples. Therefore, [66] introduced NUTS, an extension to HMC, which is tuning-free in the sense that it eliminates the need to choose the number of L and automatically sets the parameter ϵ. NUTS begins with a slice sampling step, where a slice variable u is drawn uniformly from the interval $[0, \exp(\mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p})]$, where $\boldsymbol{\theta}$ denotes the current sample and $\mathbf{p}$ is randomly drawn from $\mathcal{N}(\mathbf{p} \mid 0, \mathbf{M})$. This gives the conditional distribution $\pi(\boldsymbol{\theta}, \mathbf{p} \mid u) \sim \mathcal{U}(\boldsymbol{\theta}, \mathbf{p} \mid \{\boldsymbol{\theta}', \mathbf{p}' \mid \exp(\mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}) \geq u\})$, where $\boldsymbol{\theta}', \mathbf{p}'$ denote the proposed position and momentum respectively. Then the leapfrog integrator is used to build up a

trajectory that doubles the previous steps (either forwards or backwards) continuously. By doing this an implicit balanced binary tree is built with each leaf node corresponding to the position-momentum variables. The doubling stops when the subpath from the leftmost to the rightmost nodes of any balanced subtree of the whole binary tree starts to make a "U-turn", that is, the samples start to retrace their steps. The proposed position-momentum variables are sampled incrementally from the subtree during the doubling process, and a transition kernel that leaves the target distribution invariant is used at the end of the simulation to accept/reject the proposed new samples. In this way there is no need to choose the number of steps L. For a detailed description of the pseudo code of NUTS algorithm please refer to Algorithm 3 in [66].

3.2.2 No-U-Turn Sampler with Dual Averaging - NUTSDA

To address the issue of setting the step-size ϵ, [66] adopts an adaptation of the stochastic optimisation with a dual averaging scheme of [102], which takes the following form:

$$\epsilon_{t+1} \leftarrow \log \epsilon_1 - \frac{\sqrt{t}}{\gamma} \frac{1}{t+t_0} \sum_{i=1}^{t} H_i$$

$$\overline{\epsilon}_{t+1} \leftarrow t^{-\kappa} \log \epsilon_{t+1} + (1 - t^{\kappa}) \log \overline{\epsilon}_t$$

where t denotes the number of iterations; ϵ_1 is the initial value of epsilon, found by the heuristic that aims to obtain an acceptance rate of 0.5 using the Langevin proposal with step-size ϵ_1; $\gamma > 0$, $t_0 > 0$ are free parameters that determine the shrinkage towards $\log \epsilon_1$ and the stabilisation of the initial iterations respectively; $H_t = \delta - H^{NUTS}$, with δ denoting the desired target mean acceptance rate, H^{NUTS} being the average acceptance rate during the final iteration of doubling. The term $t^{-\kappa}(\kappa \in (0.5, 1])$ is chosen to ensure the averaged value $\overline{\epsilon}_t$ converges to a value for large t and hence the expectation of H_t (function of $\overline{\epsilon}_t$) converges to 0. For a full description of the pseudo code of NUTSDA algorithm please refer to Algorithm 6 in [66].

3.3 Metropolis-Hastings - MH

The Metropolis-Hastings algorithms [62, 95] samples from $f(\boldsymbol{\theta})$ by repeatedly considering randomly generated samples [98] and accepting the proposed moves with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}')q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{f(\boldsymbol{\theta})q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}\right\} \tag{3.4}$$

where $q(.)$ is the proposal distribution, $\boldsymbol{\theta}'$ denotes the proposed new sample, and $\boldsymbol{\theta}$ denotes the current sample. The proposal distribution $q(.)$ is commonly chosen to be a succession of random multivariate Gaussian of the form $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta}' \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\theta}$ being the former state, and covariance $\boldsymbol{\Sigma}$. Because of the symmetric property of $q(.)$, the acceptance function reduces to the form $\min\left\{1, \frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})}\right\}$. This generalises to other symmetric proposals.

Tuning the MH involves a proper choice of the covariance $\boldsymbol{\Sigma}$. However, as noted in [42], it is not trivial to select the right covariance as information about the desired target distribution is required in most cases. Very small values of $\boldsymbol{\Sigma}$ will cause slow convergence to the stationary state, whereas very large ones will lead to chains getting stuck in certain regions of the space. Ways to optimally tune the MH algorithms have been proposed in [50, 126, 127]. Approaches on adaptively tuning MH have also been reported in [58, 59, 60]. Algorithm 3 gives the pseudo code of the generic MH algorithm.

Algorithm 3 Generic MH.

Given current pair $(\boldsymbol{\theta}, f(\boldsymbol{\theta}))$

1. Draw $\boldsymbol{\theta}'$ from the proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$
2. Compute $A = \min\left\{1, \frac{f(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{f(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$
3. Draw u from $U_{[0,1]}$
4. **if** $u < A$, **return** $(\boldsymbol{\theta}', f(\boldsymbol{\theta}'))$, **else return** $(\boldsymbol{\theta}, f(\boldsymbol{\theta}))$

In cases where the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ cannot be computed analytically, it is possible to replace $f(\boldsymbol{\theta})$ with an unbiased estimate $\tilde{f}(\boldsymbol{\theta})$ to give the pseudo-marginal MH (PM-MH) transition operator [40], which will be discussed in the next section.

3.4 Pseudo-Marginal MCMC for Inference of Covariance Parameters

Standard MCMC algorithms to draw from the posterior $\pi(\boldsymbol{\theta})$ require the exact calculation of the marginal likelihood and the gradient of its logarithm. When the likelihood is not Gaussian, computing the expectation with respect to the posterior - defined in eq. (2.32) - is not feasible because of the inability to exactly calculate the marginal likelihood. In cases where the marginal likelihood can be unbiasedly estimated, it is possible to resort to pseudo-marginal MCMC approaches. Taking the Metropolis-Hastings algorithm as an example, it is possible to replace the exact calculation of the Hastings ratio

$$\frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})} = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})} \tag{3.5}$$

with an approximation where the marginal likelihood is unbiasedly estimated:

$$\frac{\tilde{f}(\boldsymbol{\theta}')}{\tilde{f}(\boldsymbol{\theta})} = \frac{\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})} \tag{3.6}$$

where $\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})$ denotes such an approximation. Interestingly, the introduction of this approximation does not affect the properties of the MCMC approach and it still yields samples from the correct posterior $\pi(\boldsymbol{\theta})$. The effect of introducing this approximation, however, is that the efficiency of the corresponding MCMC approach is reduced; this is as a result of the possibility that the algorithm accepts a proposal with a largely overestimated value of the marginal likelihood, making it difficult for any new proposals to be accepted.

By inspecting the GP marginal likelihood

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \boldsymbol{\theta})d\mathbf{f} \tag{3.7}$$

it is, therefore, possible to attempt to unbiasedly estimate this integral using importance sampling:

$$\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y} \mid \mathbf{f}_i) p(\mathbf{f}_i \mid \boldsymbol{\theta})}{q(\mathbf{f}_i \mid \mathbf{y}, \boldsymbol{\theta})} \tag{3.8}$$

Here N_{imp} is the number of samples $\mathbf{f}_i$ drawn from the importance density $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$. The motivation for attempting this approximation is to leverage the various successful attempts that construct accurate approximations to the posterior distribution over the latent variables $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \boldsymbol{\theta})$. The accuracy of the approximations to the posterior over latent variables directly affects the accuracy of the importance sampling estimates of the marginal likelihood. Despite introducing some noise in the calculation of the Hastings ratio, the resulting MCMC approach has been shown to yield state-of-the-art performance in sampling from the posterior over GP covariance parameters [40]. This thesis investigates approximations $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$ to the posterior obtained by the Laplace approximation (LA) and expectation propagation (EP) algorithms as discussed in Section 2.1.2.

3.5 Conclusion

This chapter examined the popular MCMC algorithms; that is, slice sampling, hybrid Monte Carlo including NUTS and NUTSDA, Metropolis-Hastings (MH), pseudo-marginal MCMC with an example of pseudo-marginal MH. These MCMC approaches are usually exploited to sample from the posterior over GP covariance parameters. However, here the MCMC algorithms involved repeated computations of the marginal likelihood, which are expensive in the context of Gaussian processes as discussed in Chapter 2. Accordingly, their rejection of proposals at each iteration causes a waste of expensive computations. In addition, in cases where the likelihood is non-Gaussian and hence pseudo-marginal MCMC is employed, further inefficiencies will occur because when a proposal is accepted with an overestimated marginal likelihood, it becomes difficult for the chain to accept any other proposal. Therefore, in the following chapter, alternative sampling approaches within the GP framework aiming to avoid the inefficiencies of the MCMC algorithms will be explored.

Chapter 4

Adaptive Monte Carlo

The last chapter reviewed the MCMC algorithms and analysed the inefficiencies arising from their use when sampling from the posterior over GP covariance parameters. In order to avoid such inefficiencies of the MCMC algorithms, this chapter will explore alternative inference framework for Gaussian processes based on adaptive multiple importance sampling (AMIS) [22].

Section 4.1 briefly reviews early attempts to introduce adaptation mechanism to improve the sampling efficiency of MCMC. Despite this improved efficiency, the adaptive MCMC suffers from the ergodicity issues: the resulting chain is no longer Markovian.

Section 4.2 describes the AMIS algorithm that is able to mitigate the ergodicity issues of the adaptive MCMC, with a modified version of AMIS (called MAMIS in this thesis), which aims to solve the consistency problem of AMIS, described in Section 4.2.1. This thesis is the first attempt to explore AMIS and MAMIS to sample from the posterior over GP covariance parameters.

Section 4.3 proposes the pseudo-marginal AMIS for the case of non-Gaussian likelihood where the marginal likelihood is unbiasedly estimated and never computed exactly. In this section, there is a theoretical analysis showing the unbiasedness properties of expectations computed by the proposed pseudo-marginal AMIS.

Section 4.4 concludes this chapter.

This chapter together with the following chapter (5), which reports on the experiments and results of convergence analysis of AMIS versus MCMC for both the GP regression and classification cases, are the main contributions of this thesis. The contents of these two chapters have been published in:

- X. Xiong, V. Šmídl, and M. Filippone. Adaptive multiple importance sampling for Gaussian processes. *Journal of Statistical Computation and Simulation*, 87(8):1644–1665, 2017

4.1 Introduction

Inefficiencies arising from the use of MCMC methods to sample from the posterior distribution over covariance parameters are because several proposals have been rejected. To mitigate this issue, some adaptation mechanisms of the proposals have been developed in recent years.

Two adaptation criteria are very common in the literature. One is the optimal acceptance probability, where the size and shape of the proposal distribution is scaled according to an optimal acceptance rate [50, 126]. However, as noted in [58], the hand-tuning is time consuming as the acceptance probability does not take into account the shape of the target distribution and can be difficult when the parameters are of different scales and correlated. To avoid this difficulty, other adaptation schemes employ moment matching, where moments of the proposal distribution (e.g. mean and covariance) are matched with those of the target distribution [58, 59]. A further approach to adaptation takes advantage of the regeneration of the chain [53, 130]. The work in [1] proposes a general adaptation framework using stochastic approximation schemes to learn the optimal parameters of the proposal distribution for several statistical criteria. However, devising valid adaptive MCMC methods is generally difficult in practice [1, 60].

4.2 Adaptive Multiple Importance Sampling for Gaussian Processes

The difficulty of devising adaptive MCMC approaches lies in that the chain resulting from the adaptivity is no longer Markovian, and thus more elaborate ergodicity results are needed to establish convergence to the true posterior distribution [1, 58, 59].

In response to this, Cappe et al. [14] proposed a universal adaptive sampling scheme called population Monte Carlo (PMC), where the difference from sequential Monte Carlo (SMC) [32] is that the target distribution becomes static. This method is reported to have better adaptivity than MCMC since the use of importance sampling removes the issue of ergodicity. At each iteration of PMC, the Sampling Importance Resampling (SIR) [128] particle filter is used to generate samples that are assumed to be marginally distributed from the target distribution and hence, the approach is unbiased and can be stopped at any time. Moreover, the importance distribution can be adapted using part (generated at each iteration) or all of the importance sample sequence. Douc et al. [30, 31] also introduced updating mechanisms for the weights of the mixture in D-kernel PMC which leads to a reduction either in Kullback divergence between the mixture and the target distribution or in the asymptotic variance for a function of interest. An earlier adaptive importance sampling strategy was proposed in [106].

Cornuet et al. [22] proposed a new perspective of adaptive importance sampling (AIS), called adaptive multiple importance sampling, which differs from the aforementioned PMC methods because the importance weights of all simulations, produced previously as well as currently are re-evaluated at each iteration. This method follows the 'deterministic multiple mixture' sampling scheme of [110]. The corresponding importance weight takes the form

$$w_i^t = f(\boldsymbol{\theta}_i^t) / \frac{1}{\sum_{t=0}^{T-1} N_t} \sum_{t=0}^{T-1} N_t q_t(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}}_t) \tag{4.1}$$

where T is the total number of iterations, $f(\cdot)$ denotes the target distribution $\pi(\cdot)$ up to a constant, i.e., $\pi(\cdot) \propto f(\cdot)$, $q_t(\cdot)$ denotes the importance density at iteration t with sequentially updated parameters $\widehat{\boldsymbol{\gamma}}_t$ and $\boldsymbol{\theta}_i^t$ are samples drawn from $q_t(\cdot)$ with $0 \leq t \leq T-1$, $1 \leq i \leq N_t$.

The fixed denominator in eq. (4.1) is called 'deterministic multiple mixture'. The motivation is that this construction can achieve an upper bound on the asymptotic variance of the estimator without rejecting any simulations [110]. In AMIS, the parameters $\boldsymbol{\gamma}$ of a parametric importance function $q_t(\boldsymbol{\theta};\boldsymbol{\gamma})$ are sequentially updated using the entire sequence of weighted importance samples, based on efficiency criteria such as moment matching, minimum Kullback divergence with respect to the target, or minimum variance of the weights (see, e.g. [109] for stochastic gradient-based optimisation of these efficiency criteria). This leads to a sequence of importance distributions, $q_1(\boldsymbol{\theta};\widehat{\boldsymbol{\gamma}_1}), ..., q_T(\boldsymbol{\theta};\widehat{\boldsymbol{\gamma}_T})$ that progressively improves on the approximation to the posterior over $\boldsymbol{\theta}$. Algorithm 4 gives the pseudo code of the generic AMIS algorithm.

Algorithm 4 Generic AMIS as analysed by [22].

- At iteration $t = 0$,
  1. Generate N_0 independent samples $\boldsymbol{\theta}_i^0 (1 \leq i \leq N_0)$ from the initial importance density $q_0(\boldsymbol{\theta};\widehat{\boldsymbol{\gamma}_0})$
  2. For $1 \leq i \leq N_0$, compute $\delta_i^0 = N_0 q_0(\boldsymbol{\theta}_i^0;\widehat{\boldsymbol{\gamma}_0}), w_i^0 = f(\boldsymbol{\theta}_i^0) \Big/ q_0(\boldsymbol{\theta}_i^0;\widehat{\boldsymbol{\gamma}_0})$
  3. Estimate $\widehat{\boldsymbol{\gamma}_1}$ of $q_1(\boldsymbol{\theta};\widehat{\boldsymbol{\gamma}_1})$ using the weighted samples $(\{\boldsymbol{\theta}_1^0, w_1^0\}, ..., \{\boldsymbol{\theta}_{N_0}^0, w_{N_0}^0\})$ and a well-chosen efficiency criterion for estimation.
- At iteration $t = 1, ..., T-1$,
  1. Generate N_t independent samples $\boldsymbol{\theta}_i^t (1 \leq i \leq N_t)$ from $q_t(\boldsymbol{\theta};\widehat{\boldsymbol{\gamma}_t})$
  2. For $1 \leq i \leq N_t$, compute the multiple mixture at $\boldsymbol{\theta}_i^t$

$$\delta_i^t = N_0 q_0(\boldsymbol{\theta}_i^t;\widehat{\boldsymbol{\gamma}_0}) + \sum_{l=1}^{t} N_l q_l(\boldsymbol{\theta}_i^t;\widehat{\boldsymbol{\gamma}_l})$$

  and derive the importance weights of $\boldsymbol{\theta}_i^t$

$$w_i^t = f(\boldsymbol{\theta}_i^t) \Big/ \Big[\delta_i^t \Big/ \sum_{j=0}^{t} N_j\Big]$$

3. For $0 \leq l \leq t-1$ and $1 \leq i \leq N_l$, update the past importance weights as

$$\delta_i^l \leftarrow \delta_i^l + N_t q_t(\boldsymbol{\theta}_i^l; \widehat{\boldsymbol{\gamma}}_t) \qquad \text{and} \qquad w_i^l \leftarrow f(\boldsymbol{\theta}_i^l) \Big/ \Big[\delta_i^l \Big/ \sum_{j=0}^{t} N_j\Big]$$

4. Estimate $\widehat{\boldsymbol{\gamma}_{t+1}}$ using all the weighted samples

$$(\{\boldsymbol{\theta}_1^0, w_1^0\}, ..., \{\boldsymbol{\theta}_{N_0}^0, w_{N_0}^0\}, ..., \{\boldsymbol{\theta}_1^t, w_1^t\}, ..., \{\boldsymbol{\theta}_{N_t}^t, w_{N_t}^t\})$$

and the same efficiency criterion for estimation.

This thesis has used a Gaussian importance density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\gamma}_t = (\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$. Moment matching has been chosen as the efficiency criterion to estimate $\widehat{\boldsymbol{\gamma}}_t = (\hat{\boldsymbol{\mu}}^t, \hat{\boldsymbol{\Sigma}}^t)$ using the self-normalised AMIS estimator:

$$\hat{\boldsymbol{\mu}}^t = \frac{\sum_{l=0}^{t}\sum_{i=1}^{N_l} w_i^l \boldsymbol{\theta}_i^l}{\sum_{l=0}^{t}\sum_{i=1}^{N_l} w_i^l} \qquad \text{and} \qquad \hat{\boldsymbol{\Sigma}}^t = \frac{\sum_{l=0}^{t}\sum_{i=1}^{N_l} w_i^l (\boldsymbol{\theta}_i^l - \hat{\boldsymbol{\mu}}^t)(\boldsymbol{\theta}_i^l - \hat{\boldsymbol{\mu}}^t)^T}{\sum_{l=0}^{t}\sum_{i=1}^{N_l} w_i^l}$$

4.2.1 Modified AMIS - MAMIS

Despite the efficiency brought by AMIS compared with other AIS techniques, proving convergence of this algorithm is not straightforward. The work in [92] proposed a modified version of AMIS called MAMIS, aiming at obtaining a variant of AMIS where convergence can be more easily established. The difference is that the new parameters $\widehat{\boldsymbol{\gamma}}_t$ are estimated based only on samples produced at iteration t, i.e. $\boldsymbol{\theta}_1^t, ..., \boldsymbol{\theta}_{N_t}^t$, with classical weights $f(\boldsymbol{\theta}_i^t)/q(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}}_t)$. Then the weights of all simulations are updated according to eq. (4.1) to give the final output. The sample size N_t is suggested to grow at each iteration so as to improve the accuracy of $\widehat{\boldsymbol{\gamma}}_t$. MAMIS effectively solves any convergence issues of AMIS, but convergence is slower because fewer samples are used to update the importance distribution. Algorithm 5 describes modified AMIS.

Algorithm 5 MAMIS as analysed by [92].

Given an initial importance density $q_1(\boldsymbol{\theta}; \widehat{\boldsymbol{\gamma}_1})$ and increasing sample sizes N_1, ..., N_T.

- For $1 \leq t \leq T$,

  1. Generate N_t independent samples $\boldsymbol{\theta}_i^t (1 \leq i \leq N_t)$ from $q_t(\boldsymbol{\theta}; \widehat{\boldsymbol{\gamma}}_t)$
  2. For $1 \leq i \leq N_t$, compute the importance weights of $\boldsymbol{\theta}_i^t$

$$w_i^t = f(\boldsymbol{\theta}_i^t) \Big/ q_t(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}}_t)$$

  3. Estimate $\widehat{\boldsymbol{\gamma}_{t+1}}$ using the weighted samples

$$(\{\boldsymbol{\theta}_1^t, w_1^t\}, ..., \{\boldsymbol{\theta}_{N_t}^t, w_{N_t}^t\})$$

     and a well-chosen efficiency criterion for estimation.

- For $1 \leq t \leq T$ and $1 \leq i \leq N_t$, update the weights of all the simulations

$$w_i^t = f(\boldsymbol{\theta}_i^t) / \frac{1}{\sum_{t=1}^T N_t} \sum_{t=1}^T N_t q_t(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}}_t)$$

  and return the weighted samples $(\{\boldsymbol{\theta}_1^1, w_1^1\}, ..., \{\boldsymbol{\theta}_{N_1}^1, w_{N_1}^1\}, ..., \{\boldsymbol{\theta}_1^T, w_1^T\}, ..., \{\boldsymbol{\theta}_{N_T}^T, w_{N_T}^T\})$.

For the illustration of MAMIS in this thesis, a Gaussian importance density and the moment matching criterion have been used similar to those used for AMIS in Section 4.2.

4.3 Pseudo-Marginal AMIS

The above AMIS/MAMIS estimators are designed for the general analytically computable marginal likelihood, such as in the case of GP regression. In this thesis, it is proposed to use AMIS to sample from the posterior over model parameters where the likelihood is analytically intractable but can be unbiasedly estimated. In practice, we modify AMIS by

replacing the exact calculation of the marginal likelihood with an unbiased estimate, giving an unbiased estimate of the posterior up to a normalising constant:

$$\tilde{f}(\boldsymbol{\theta}) = \tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4.2}$$

This is referred to as pseudo-marginal AMIS (PM-AMIS), inspired by the name pseudo-marginal MCMC that was given to the class of MCMC algorithms replacing exact calculations of the likelihood with unbiased estimates [2]. The pseudo-code of PM-AMIS is similar to that of AMIS described in Algorithm 4, except that the target distribution up to a constant $f(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ is replaced by the above unbiased estimate $\tilde{f}(\boldsymbol{\theta})$.

Despite the fact that calculations are approximate, pseudo-marginal MCMC methods yield samples from the correct posterior distribution over covariance parameters, so a natural question is whether the same argument holds for this proposal. The remainder of this section provides an analysis of the properties of pseudo-marginal AMIS, discussing the conditions under which it yields unbiased expectations with respect to the posterior distribution over covariance parameters. As in [114, 145], a random variable z is introduced whose distribution (denoted by $p(z \mid \boldsymbol{\theta})$ herein) is determined by the randomness occurring when carrying out the unbiased estimation of the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. Define:

$$z = \log \tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) - \log p(\mathbf{y} \mid \boldsymbol{\theta}) \tag{4.3}$$

i.e.

$$\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})e^{z} \tag{4.4}$$

Because of the unbiased property ($E[\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})] = p(\mathbf{y} \mid \boldsymbol{\theta})$), it is possible to readily verify that $E[e^{z}] = 1$. For the sake of clarity, it is useful to define the unnormalised joint density of z and $\boldsymbol{\theta}$ as:

$$f(z, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})e^{z}p(z \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4.5}$$

with a corresponding normalised version

$$\pi(z, \boldsymbol{\theta}) = \frac{f(z, \boldsymbol{\theta})}{Z} \tag{4.6}$$

Marginalising this joint density with respect to z

$$\int \pi(z, \boldsymbol{\theta}) dz = \int \frac{f(z, \boldsymbol{\theta})}{Z} dz = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{Z} E[e^z] = \frac{f(\boldsymbol{\theta})}{Z} \tag{4.7}$$

yields the target posterior $\pi(\boldsymbol{\theta})$ of interest defined in eq. (2.31).

Recall that the objective is analysing expectations under the posterior over the parameters $\pi(\boldsymbol{\theta})$ of some function $h(\boldsymbol{\theta})$

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int h(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} \tag{4.8}$$

The analysis started by substituting eq. (4.7) into eq. (4.8), obtaining

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) \frac{f(z, \boldsymbol{\theta})}{Z} d\boldsymbol{\theta} dz \tag{4.9}$$

In PM-AMIS, let N_t denote the number of samples generated at each iteration t, $q_t(\boldsymbol{\theta})$ denote the importance density at each iteration for $\pi(\boldsymbol{\theta})$. We also define

$$q_t(z, \boldsymbol{\theta}) = p(z \mid \boldsymbol{\theta}) q_t(\boldsymbol{\theta}) \tag{4.10}$$

as the joint importance density at each iteration for $\pi(z, \boldsymbol{\theta})$, $(z_i^t, \boldsymbol{\theta}_i^t)$ as samples drawn from $q_t(z, \boldsymbol{\theta})$ with $0 \leq t \leq T, 1 \leq i \leq N_t$.

Since in a practical setting $f(z, \boldsymbol{\theta})$ is the only function that we can evaluate, the expectation defined in eq. (4.9) is estimated by the self-normalised PM-AMIS estimator:

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] \approx \frac{1}{\sum_{t=0}^{T} \sum_{i=1}^{N_t} {w_i}^t} \sum_{t=0}^{T} \sum_{i=1}^{N_t} {w_i}^t h(\boldsymbol{\theta}_i^t) \tag{4.11}$$

where the weights of this estimator are computed as

$${w_i}^t = \frac{f(z_i^t, \boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l q_l(z_i^t, \boldsymbol{\theta}_i^t)} \tag{4.12}$$

Expanding the terms in the computations of the weights, namely substituting eq. (4.5) and

eq. (4.10) into eq. (4.12) gives:

$$\begin{aligned} {w_i}^t &= \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_i^t) e^{z_i^t} p(z_i^t \mid \boldsymbol{\theta}_i^t) p(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l p(z_i^t \mid \boldsymbol{\theta}_i^t) q_l(\boldsymbol{\theta}_i^t)} \\ &= \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_i^t) e^{z_i^t} p(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l q_l(\boldsymbol{\theta}_i^t)} \end{aligned} \tag{4.13}$$

which can be rewritten in terms of the unbiased estimate of the marginal likelihood as:

$${w_i}^t = \frac{\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}_i^t) p(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l q_l(\boldsymbol{\theta}_i^t)} = \frac{\tilde{f}(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l q_l(\boldsymbol{\theta}_i^t)} \tag{4.14}$$

Equation (4.14) shows how the importance weights can be computed by the unbiased estimator of the marginal likelihood.

It is now time to appeal to Lemma 1 in [22], which gives the conditions under which the self-normalised estimator of AMIS will converge to eq. (4.8). Following the conditions in Lemma 1 in [22], when T and N_0, ..., N_{T-1} are fixed, and when N_T goes to infinity, ${w_i}^t$ (eq. (4.12)) becomes:

$${w_i}^t \simeq \frac{f(z_i^t, \boldsymbol{\theta}_i^t)}{q_T(z_i^t, \boldsymbol{\theta}_i^t)} \tag{4.15}$$

Then we have

$$\begin{aligned} E_{q_t(z,\boldsymbol{\theta})} \left[\frac{1}{\sum_{t=0}^{T} \sum_{i=1}^{N_t} {w_i}^t} \sum_{t=0}^{T} \sum_{i=1}^{N_t} {w_i}^t h(\boldsymbol{\theta}_i^t) \right] &= \frac{1}{Z \sum_{t=0}^{T} N_t} \sum_{t=0}^{T} N_t \int h(\boldsymbol{\theta}) \frac{f(z,\boldsymbol{\theta})}{q_T(z,\boldsymbol{\theta})} q_T(z,\boldsymbol{\theta}) d\boldsymbol{\theta} dz \\ &= \frac{1}{\sum_{t=0}^{T} N_t} \sum_{t=0}^{T} N_t \int h(\boldsymbol{\theta}) \frac{f(z,\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} dz \\ &= \frac{1}{\sum_{t=0}^{T} N_t} \sum_{t=0}^{T} N_t \int h(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} \\ &= \int h(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} = E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] \end{aligned}$$

where the normalising constant Z is estimated by $\frac{\sum_{t=0}^{T} \sum_{i=1}^{N_t} {w_i}^t}{\sum_{t=0}^{T} N_t}$.

Therefore, under the conditions that T and N_0, ..., N_{T-1} are fixed and that N_T goes to infinity, which are the same conditions mentioned in Lemma 1 in [22], the estimator of eq. (4.11) proves to be an unbiased estimator of $E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})]$. As noted in [22], these conditions might

prove restrictive in practice; however, these conditions provide some solid grounds onto which convergence can be established for AMIS. Furthermore, in a practical setting, when in doubt as to whether convergence might be an issue, it is always possible to switch to the modified version of AMIS [92] during execution.

4.4 Conclusion

In this chapter, an alternative inference framework based on AMIS to sample from the posterior over GP covariance parameters was proposed in order to mitigate the effect of the inefficiencies of MCMC methods. In the case of a Gaussian likelihood, AMIS has been proposed for computing expectations under the posterior over GP covariance parameters. In the case of non-Gaussian likelihoods, pseudo-marginal AMIS that extends AMIS to deal with GP models where the marginal likelihood cannot be computed exactly and hence is unbiasedly estimated has been proposed, and the unbiased property of expectations computed by the proposed Pseudo-Marginal AMIS has been theoretically proved.

The next chapter will empirically examine the sampling efficiency of AMIS versus MCMC for GP models, where the sampling efficiency is measured in terms of convergence speed against computational complexity.

Chapter 5

Experiments and Results

This chapter sees a comparison of the state-of-the-art MCMC methods for GP models against the proposed AMIS. The comparison targets sampling efficiency (in terms of convergence speed) versus computational complexity. The aim of the experiments is to discover whether adaptive importance sampling (AMIS/MAMIS) can improve the speed of convergence with respect to computational complexity compared to MCMC approaches.

The rest of this chapter is organised as follows. Section 5.1 summarises the tuning parameters of the competing MCMC approaches. Section 5.2 presents the data sets used in the experiments. Section 5.3 describes the experimental setup, with settings for GP regression and classification given in Section 5.3.1 and Section 5.3.2 respectively. Section 5.4 proposes a new metric for the comparison of the convergence analysis of MCMC versus AMIS for GP models. Section 5.5 presents the experimental results, and Section 5.6 offers a conclusion.

5.1 Competing Sampling Methods

This section summarises the turning parameters of the MCMC approaches (a survey of which can be found in Chapter 3). Table 5.1 gives the tuning parameters of the competing sampling algorithms.

Table 5.1 Competing sampling algorithms.

Sampler	**Tuning parameters**
Metropolis-Hastings (MH)	Covariance matrix $\boldsymbol{\Sigma}$
Hybrid Monte Carlo (HMC)	Mass matrix $\boldsymbol{\Sigma}$, Leapfrog stepsize ϵ, Number of leapfrog steps L
No-U-Turn Sampler (NUTS)	Mass matrix $\boldsymbol{\Sigma}$, Leapfrog stepsize ϵ
NUTS with Dual Averaging (NUTSDA)	Mass matrix $\boldsymbol{\Sigma}$
Slice Sampling (SS)	Width of the initial bracket

5.2 Data Sets

The sampling methods considered in this work were tested on six benchmark data sets from the University of California, Irvine (UCI) repository [4]. The Concrete, Housing and Parkinsons data sets are for GP regression, whereas the Glass, Thyroid and Breast data sets are for GP classification. The number of data points and features for each data set are given in Table 5.2. For the original Parkinsons data set we randomly sampled 4 records for each of the 42 patients, resulting in 168 data points in total.

Table 5.2 Data sets

	Data sets for regression			Data sets for classification		
	Concrete	Housing	Parkinsons	Glass	Thyroid	Breast
n	1030	506	168	214	215	682
d	8	13	20	9	5	9

n denotes the number of data points
d denotes the number of features.

5.3 Experimental Setup

This section presents the experimental settings of the MCMC and AMIS/MAMIS samplers for both the GP regression and classification cases, given in Section 5.3.1 and Section 5.3.2, respectively.

5.3.1 Settings for GP Regression

Three different covariances have been compared for the proposals of the MH algorithm. The first is proportional to the identity matrix. The second and third covariances are proportional to the inverse of the negative Hessian of the log-posterior (denoted by $\mathbf{H}$) evaluated at the mode (denoted by $\mathbf{m}$); one uses the full Hessian matrix, whereas the other uses its diagonal only, namely $\mathrm{diag}((-\mathbf{H})^{-1})$. The mode $\mathbf{m}$ is found by the maximum likelihood optimisation using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.

Thus the proposals that are compared in this work take the form of $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha\mathbf{I})$, $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha(-\mathbf{H})^{-1})$, and $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha\,\mathrm{diag}((-\mathbf{H})^{-1}))$, where α is a tuning parameter. The parameter α is tuned in pilot runs until the desired acceptance rate (approximately 25%) is reached, as suggested by [125].

The approximate distribution $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, (-\mathbf{H})^{-1})$ is used as the initial importance density for AMIS/MAMIS. This approximation is also used to initialise several other samplers considered in this work (listed in Table 5.1). In this way, valid summary inference from multiple independent sequences can be obtained [51]. For AMIS/MAMIS, two different strategies were explored to update the covariance of the importance density. One updates the full covariance, whereas the other updates the diagonal of the covariance only. The first two rows of Table 5.3 show the experimental settings for AMIS/MAMIS.

Motivated by the knowledge of the scales and the correlations of the position variables can improve the performance of HMC [101], three types of mass matrices for HMC were chosen, namely the identity matrix, the inverse of the approximate covariance, and the inverse of the diagonal of the approximate covariance. The maximum leapfrog step was set to be 10. The stepsize ϵ was tuned until a suggested acceptance rate (approximately 65%) is reached [9, 101]. The three forms of mass matrix apply to NUTS, NUTSDA as well; a full description of the pseudo codes of these algorithms can be found in Algorithms 3 and 6 in [66], respectively. NUTS requires the tuning of a stepsize ϵ. After a few trials, the stepsize of NUTS was set to 0.1. Although tuning leapfrog steps and stepsize is not an issue in NUTSDA, the parameters (γ, t_0, κ) for the dual averaging scheme therein have to be tuned by hand to produce reasonable results. After trying a few settings for each parameter, it was

decided to proceed with the values $\gamma = 0.05$, $t_0 = 30$, and $\kappa = 0.75$ in both the RBF and ARD covariance cases.

The slice sampling algorithm adopted in this thesis makes component-wise updates of the parameters, where a new sample was drawn according to the 'stepping out' and 'shrinkage' procedures as described in [100]. In these implementations, the estimate of the typical size of a slice w was set to 1.5.

Table 5.3 Settings for AMIS/MAMIS/PM-AMIS.

	RBF covariance		ARD covariance	
	T	N_t	T	N_t
AMIS	1120	25	280	100
MAMIS	46	$26t$	5	$3000 + 1000t$
PM-AMIS	60	400	60	400

T is the total number of iterations.
N_t is the sample size at each iteration t.

5.3.2 Settings for GP Classification

As a representative example of GP models with non-Gaussian likelihoods, a probit classification was considered. Since the likelihood is analytically intractable and thus unbiasedly estimated, the critical property of reversibility and preservation of volume of HMC, NUTS, and NUTSDA is no longer satisfied. In addition, slice sampling with the noisy estimate $\tilde{f}(\boldsymbol{\theta})$ is still valid, but naively running standard SS with the noisy estimate $\tilde{f}(\boldsymbol{\theta})$ worked very poorly as reported in [97]. As a result, only PM-AMIS and pseudo-marginal MH (PM-MH) have been compared to infer covariance parameters in GP classification.

Both the EP and LA approximations are used to obtain importance densities to unbiasedly estimate the marginal likelihood. The last row of Table 5.3 shows the settings of PM-AMIS in both the RBF and ARD cases except for the Breast data set in the ARD case using LA approximation, where the total number of iterations T was set to 240 for the sake of presentation. The initial importance density is obtained by the same optimisation method as described in Section 5.3.1 except that the gradient required to perform the optimisation cannot be computed analytically but is estimated from the EP or LA approximations. The full

covariance of the importance density is updated during the adaptation process. The proposal of PM-MH also takes the form of $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha(-\mathbf{H})^{-1})$ where H is the Hessian matrix obtained again from the EP or LA approximate marginal likelihood. The collection of samples follows an initial tuning of α to reach the recommended acceptance rate of approximately 25%.

5.4 Convergence Analysis

Since the classic $\hat{R}$ score is for MCMC convergence analysis and is not suitable for importance sampling, convergence analysis here is performed based on the IQR (interquartile range) of the expectation of norm of parameters ($E_{p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X})}[\,\|\boldsymbol{\theta}\|]$) over several repetitions against the number of $O(n^3)$ operations. This is reported to be a more reliable measure of complexity than running time, as many other factors, which do not relate directly to the actual computing complexity of the algorithm, can affect the running time [42]. In GP regression the IQR is computed over 100 repetitions, whereas in GP classification it is based on 20 repetitions.

For AMIS/MAMIS/SS/MH, the computational complexity lies in the computation of the function of $f(\boldsymbol{\theta})$, where one $O(n^3)$ operation is required to perform the Cholesky decomposition of the covariance matrix $\mathbf{C}$. For HMC/NUTS/NUTSDA where computing the gradient is necessary, two extra $O(n^3)$ operations are needed for the computation of the inverse of the covariance matrix $\mathbf{C}$.

For PM-AMIS/PM-MH, the computational complexity largely comes from the EP or LA approximation of the posterior of the latent variables in order to compute the unbiased estimate $\tilde{f}(\boldsymbol{\theta})$. Both EP and LA approximations require two Cholesky decomposition ($O(n^3)$ operations); one is for the decomposition of the covariance matrix $\mathbf{K}$ of the GP prior, while the other is for the decomposition of the covariance of the approximating Gaussian. Each iteration of EP and LA requires five $O(n^3)$ operations and one $O(n^3)$ operation, respectively. In the LA approximation, two extra $O(n^3)$ operations are needed to compute the covariance of the Gaussian approximation.

5.5 Results

This section reports on the experimental results regarding an extensive comparison of MCMC versus AMIS for GP models in terms of sampling efficiency, measured by the convergence speed against computational complexity. Convergence analyses of samplers for GP regression and classification are described in Section 5.5.1 and Section 5.5.2, respectively.

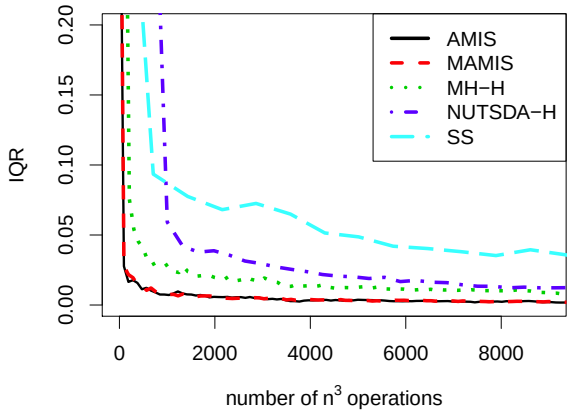
Table 5.4 Notation for the samplers used in the experiments.

AMIS/MAMIS	AMIS/MAMIS for GP regression where the full covariance matrix of the proposal distribution is updated at each iteration
AMIS-D/MAMIS-D	AMIS/MAMIS for GP regression where only the diagonal of the covariance matrix of the proposal distribution is updated at each iteration
MH-I	MH for GP regression where the covariance of the starting proposal distribution for tuning is the identity matrix
MH-D	MH for GP regression where the covariance of the starting proposal distribution for tuning is the diagonal of the approximate covariance from the optimisation
MH-H	MH for GP regression where the covariance of the starting proposal distribution for tuning is the approximate covariance from the optimisation
HMC-I/NUTS-I/NUTSDA-I	HMC family for GP regression where the mass matrix is the identity matrix
HMC-D/NUTS-D/NUTSDA-D	HMC family for GP regression where the mass matrix is the inverse of the diagonal of the approximate covariance from the optimisation
HMC-H/NUTS-H/NUTSDA-H	HMC family for GP regression where the mass matrix is the inverse of the approximate covariance from the optimisation
PM-AMIS	AMIS for GP classification where the full covariance matrix of the proposal distribution is updated at each iteration
PM-MH	MH for GP classification where the covariance of the starting proposal distribution for tuning is the approximate covariance from the optimisation

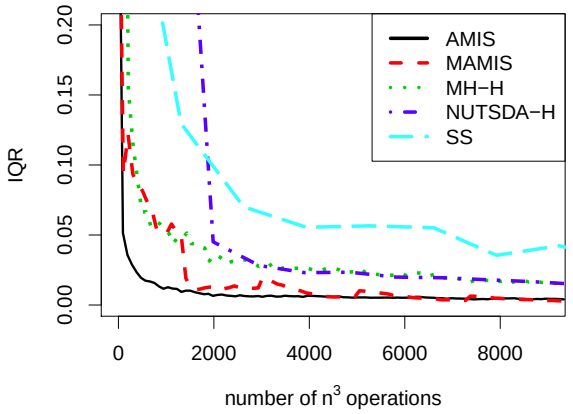
5.5.1 Convergence of Samplers for GP Regression

This section presents the comparison of convergence of samplers for GP regression considered in the experiments (Table 5.4). Details of convergence results of AMIS family (AMIS/MAMIS and their variants), MH family (MH-I/MH-D/MH-H) and HMC family (standard HMC , NUTS, NUTSDA) for the three regression data sets can be found in Appendix A.

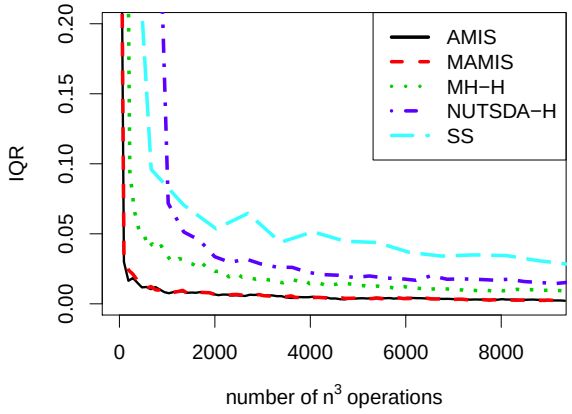
Figure 5.1 shows the results of AMIS compared to the various competitors, where for the sake of brevity, only the results of their best configurations are reported. The results are shown for the three regression data sets for both the RBF (Figures 5.1(a), 5.1(c), 5.1(e)) and ARD (Figures 5.1(b), 5.1(d), 5.1(f)) covariances. It is interesting to see that AMIS/MAMIS performs best among all methods in terms of convergence speed in the RBF covariance case. In the ARD covariance case, AMIS also converges much faster than the other approaches. However, these experiments show that in this case, although MAMIS converges faster than the other approaches in the Concrete data set, it converges slowly in the Housing and Parkinsons data sets, which is possibly because of the higher dimensionality compared to the previous cases.



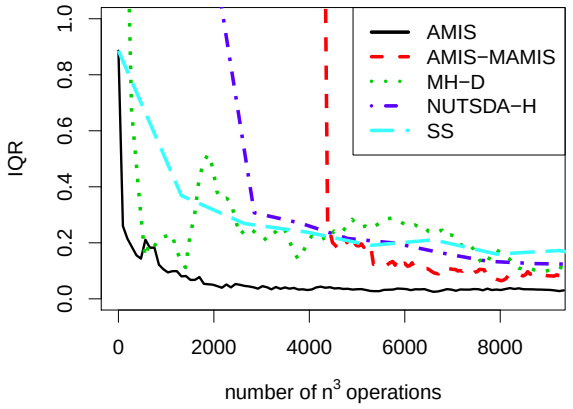
(a) Concrete data set - RBF



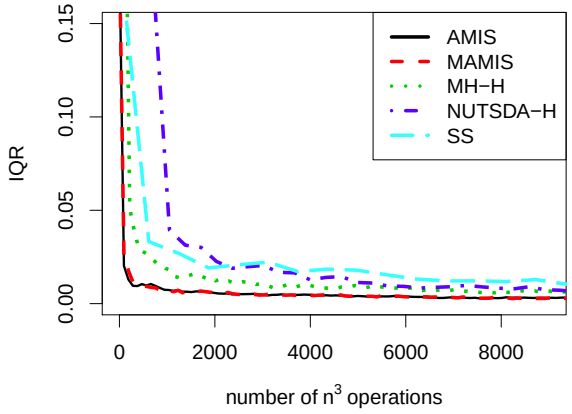
(b) Concrete data set - ARD



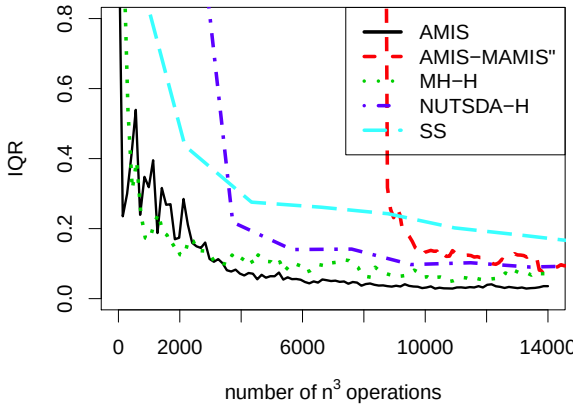
(c) Housing data set - RBF



(d) Housing data set - ARD



(e) Parkinsons data set - RBF



(f) Parkinsons data set - ARD

Figure 5.1 Convergence of AMIS, Best of MAMIS, Best of MH family, Best of HMC family, SS for GP regression.

In cases where MAMIS converges slowly, the fact that AMIS converges faster can be exploited to construct hybrid sampling schemes where MAMIS is initialised from a run of AMIS. In this way, it is possible to leverage the fast adaptation of AMIS, while ensuring that the overall scheme does not introduce any bias. In the experiments, this AMIS-MAMIS combination was tested in cases where MAMIS converges slowly. These results are reported in Figure A.5(f), A.6(f) where EOT (end of tuning) indicates the point where there was a switch to MAMIS. Three settings (Table 5.5) of AMIS-MAMIS were tested for the Parkinsons data set.

Table 5.5 Settings for AMIS-MAMIS

	N_t for MAMIS	$\star$ number of tuning samples for MAMIS	$\star\star$ corresponding tuning cost
AMIS-MAMIS	$1000t$	13000	4333
AMIS-MAMIS'	$5000t$	13000	4333
AMIS-MAMIS"	$5000t$	26000	8667

N_t is the sample size at each iteration t.
$\star$ This refers to the number of samples generated from AMIS for tuning the initial importance density of MAMIS.
$\star\star$ Unit of the tuning cost: number of n^3 operations.

For the Housing data set, only AMIS-MAMIS in Table 5.5 was tested. The results for the Housing and Parkinsons data sets in the ARD covariance case prove the convergence of AMIS-MAMIS. In particular, AMIS-MAMIS and AMIS-MAMIS" seem to compete well with the other MCMC approaches in terms of convergence for the Housing data set and the Parkinsons data set respectively. As shown in Figure A.6(f), the best performance of AMIS-MAMIS" for the Parkinsons data set suggests that for higher dimensional problems, a more accurate initialisation and a larger sample size at each iteration for MAMIS are necessary to achieve faster convergence.

Another attempt made in this thesis to improve convergence speed of the adaptive importance sampling schemes was to regularise the estimation of the parameters of the importance distribution as illustrated in [161]. This regularisation stems from the use of an informative prior on $\boldsymbol{\gamma}$ of the importance distribution $q_t(\boldsymbol{\gamma})$ of MAMIS and treats the update of these parameters in a Bayesian fashion [79]. This construction makes it possible to avoid situations where

the importance distribution degenerates to low rank as a result of few importance weights dominating the rest. In this work, an informative prior based on a Gaussian approximation to the posterior over covariance parameters has been used. This method has been denoted by MAMIS-P and in the ARD covariance case it was tested only on the Housing data set. The result indicates that even though MAMIS-P improves on MAMIS, its convergence is slower than AMIS-MAMIS (Figure A.5(f)).

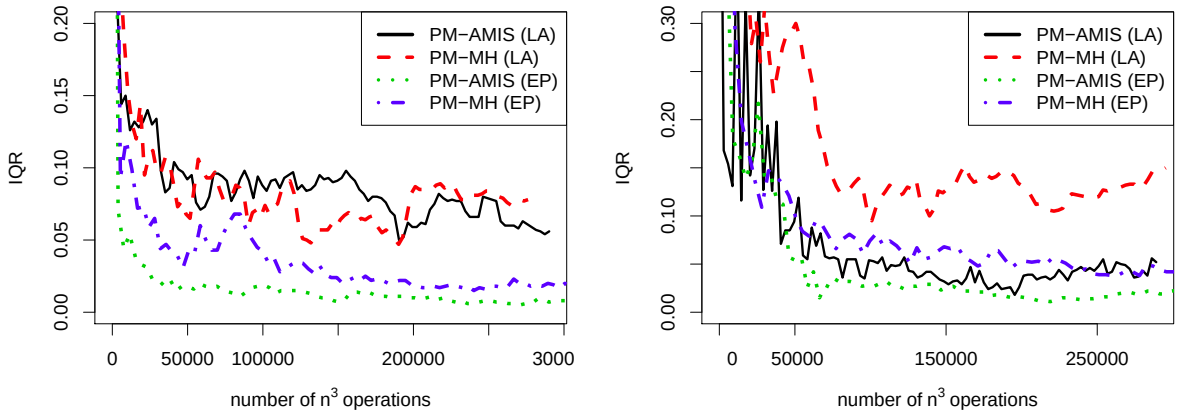
5.5.2 Convergence of Samplers for GP Classification

The comparison of convergence of PM-AMIS and PM-MH for GP classification is presented in this section.

Figure 5.2 shows the convergence results of PM-AMIS/PM-MH using EP and LA approximation with $N_{imp} = 64$, where N_{imp} denotes the number of importance samples of latent variables $\mathbf{f}$ to estimate the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. Figures 5.2 (a), 5.2 (c), 5.2 (e) are the results for the RBF covariance case, whilst Figures 5.2 (b), 5.2 (d), 5.2 (f) display the results for the ARD covariance case. In these figures, EP represents the case where the Gaussian approximation to the posterior of latent variables $\mathbf{f}$ is obtained by EP approximation, whereas LA denotes the case where the Gaussian approximation is obtained by LA approximation.

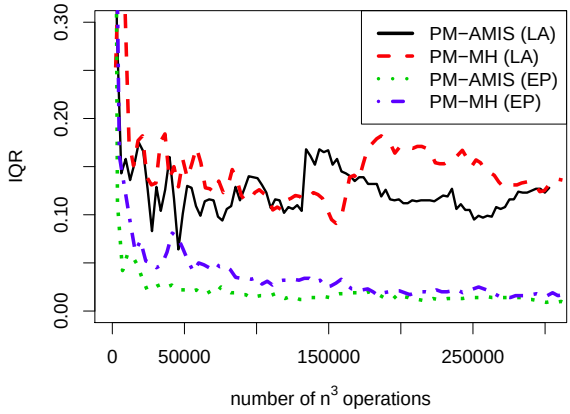
The results indicate that PM-AMIS is competitive with PM-MH in terms of convergence speed in all the EP approximation cases and in most of the LA approximation cases. The results also seem to suggest that PM-AMIS/PM-MH converge faster with the EP approximation than with the LA approximation in most cases, which can be attributed to the fact that EP yields a more accurate approximation to the posterior over covariance parameters than LA [80, 104].

The performance of PM-AMIS and PM-MH with $N_{\text{imp}} = 1$ were also tested, the results of which are shown in Appendix B. As expected, both PM-AMIS and PM-MH algorithms with a higher number of importance samples converge much faster than those with a lower number of importance samples.

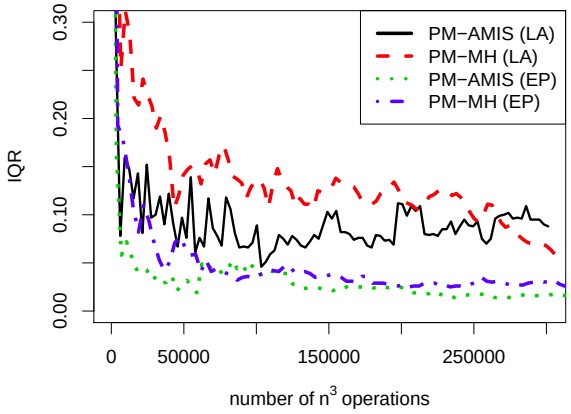


(a) Glass data set - RBF

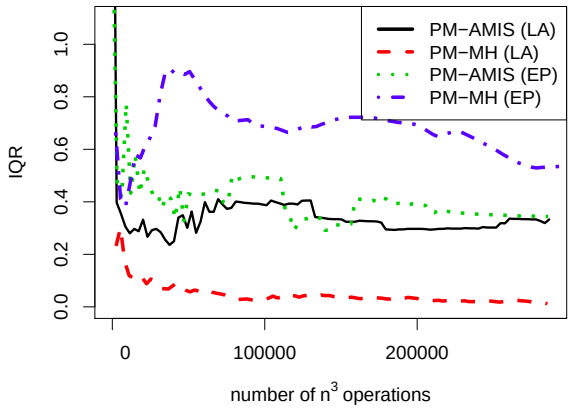
(b) Glass data set - ARD



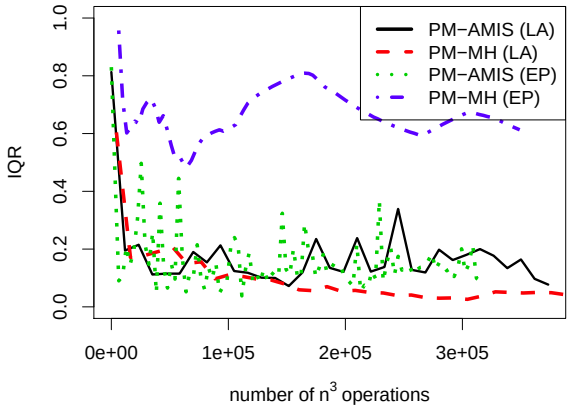
(c) Thyroid data set - RBF



(d) Thyroid data set - ARD



(e) Breast data set - RBF



(f) Breast data set - ARD

Figure 5.2 Convergence of Best of PM-AMIS, Best of PM-MH using EP and LA approximation for GP classification. LA in the brackets indicates the case where the Gaussian approximation to the posterior of the latent variables used in the corresponding method is obtained by LA approximation, whereas EP in the brackets indicates the case where the Gaussian approximation is obtained by EP approximation.

5.6 Conclusion

This thesis has proposed the use of adaptive importance sampling techniques to compute expectations under the posterior distribution of covariance parameters in Gaussian processes. The motivation for this proposal was based on a number of observations related to the complexity of dealing with the calculation of the marginal likelihood. In GPs with a Gaussian likelihood, calculating the marginal likelihood and the gradient of its logarithm with respect to covariance parameters is expensive and the rejection of proposals of standard MCMC algorithms leads to a waste of computations. In GPs with non-Gaussian likelihoods, pseudo marginal MCMC approaches bypass the need to compute the marginal likelihood exactly, but may suffer from inefficiencies because when a proposal is accepted and the marginal likelihood is largely overestimated, it becomes difficult for the chain to accept any other proposal. A further motivation behind this work is that importance sampling-based algorithms are generally easy to implement and tune, and can be massively parallelised.

The extensive set of results reported in this chapter supports the intuition that importance sampling-based inference of covariance parameters is competitive with MCMC algorithms. In particular, the results indicated that it is possible to achieve convergence of expectations under the posterior distribution of covariance parameters faster than employing MCMC methods in a wide range of scenarios. Even in the case of around twenty parameters, where importance sampling-based methods start to degrade in performance, this proposal is still competitive with MCMC approaches.

The following chapter will examine the application of the proposed PM-AMIS in the area of social signal processing, in particular for classifying the personality traits of individuals based on their favourite pictures posted on a photo-sharing platform.

Chapter 6

Gaussian Processes for Finding Difference Makers in Personality Impressions - an Application of PM-AMIS

Flickr (a popular photo-sharing platform) allows its users to generate galleries of "faves", i.e. pictures that they have tagged as favourite. According to recent studies, the faves are predictive of the personality traits that people attribute to Flickr users. This chapter investigates this phenomenon. The experiments were performed over the PsychoFlickr Corpus - $60,000$ pictures tagged as favourite by 300 Flickr users. The experimental results of this chapter showed that faves can be used to predict whether a Flickr user is perceived to be above median or not with respect to each of the Big-Five personality traits [132]. The accuracies range between 58% and 79% depending on the particular trait. The task has been performed with the PM-AMIS classifier based on Gaussian processes (proposed in Section 4.3), together with a newly designed kernel - the *Group Automatic Relevance Determination* (G-ARD) kernel.

The reasons for choosing the GP-based PM-AMIS classifier are as follows. Apart from being a non-parametric and fully probabilistic approach, which is capable of capturing more uncertainty in the data, the main novelty of the approach is the G-ARD kernel. Its accuracies are comparable, if not superior, to those achieved with SVM, a widely applied state-of-the-

art classifier. However, the most important advantage of the G-ARD is that its parameter set includes weights - set automatically during the training process - capable of identifying the feature groups that better account for the classification outcome. In this respect, the G-ARD has been inspired by the *Automatic Relevance Determination* (ARD) kernel [90]. The main difference is that this latter has a weight for each individual feature and, therefore, the number of its parameters tends to be larger. Therefore, the G-ARD appears to be more suitable when the amount of training material is limited like this case.

In these experiments, the analysis of the G-ARD weights shows that the classification outcome depends, to a significant extent, on the following characteristics of the faves: presence of human faces, composition, textural properties and, finally, number and size of visually homogeneous regions. Furthermore, weight differences across personality assessors of different national origin provide indications about cultural effects.

The prediction of personality traits has been addressed extensively in the literature (see [157] for an extensive survey). However, this is the first work that has tried to go beyond the simple classification to provide insights about the actual interplay between the data and the attributed traits. This applies, in particular, to previous results obtained over the publicly available data used in this work [24, 134].

The results in this chapter were published in:

- X. Xiong, M. Filippone, and A. Vinciarelli. Looking good with Flickr Faves: Gaussian processes for finding difference makers in personality impressions. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 412–415, 2016

The rest of this chapter is organised as follows: Section 6.1 shows the importance of investigating the interplay between the trace people leave on social media and the impressions these traces convey, Section 6.2 presents state-of-the-art in personality inference from social media data, Section 6.3 describes the data and the personality model adopted in this work, Section 6.4 presents the features, Section 6.5 introduces the G-ARD approach, Section 6.6 reports on the experimental results and, finally, Section 6.7 draws some conclusions.

6.1 Introduction

Photos and videos are seen as key "*social currencies*" [118] online because of their ubiquitous presence. According to [118], 56% of the American internet users either post self-made pictures or videos online, or re-post those they have found online. With the ease and convenience of digital technologies, such as mobile broadband (including wifi) as well as smartphones with built-in cameras, it is easy for people to take and share pictures or videos (e.g. life chronicling) online. The work of [34] has shown that 82% of the American users use mobile phones to take pictures while 56% of the users use them to access the Internet.

Personal photography is described in terms of its social uses as multiple, overlapping technologies of personal and group memory, relationships, self-representation and expressiveness [68]. According to [137], "*[...] the image is an extension of one's identity, reflecting aspects of one's personality, relationships, and lifestyle [...]. In some cases, the image gives expression to the unconscious dimensions of one's character*". In addition, photographic images (self-portraits, pictures of friends, events and so forth) can both demonstrate the photographer's skill or aesthetic sense and reflect his/her unique point of view or creativity, and thus are often seen as an alternative to direct forms of interaction such as emails [67]. This "*pictorial communication*", [17] that provides social and emotional support, information resources and connections to other people [163], has also created the widespread popularity of online photo-sharing communities such as Flickr.

Furthermore, in many social media platforms (e.g. Flickr, Facebook and Twitter), the pervasive "*liking*" mechanism - the fav/like button underneath an item (e.g. photo and shared link) by clicking on which viewers mark that item as a favourite - is reported to have acquired several social and psychological functions (emotional impact, artistic merit, social support and social barter) [137]. The significant role of such like/favourite actions in interpersonal interactions is also evidenced by the fact that "*likes*" happen most frequently among connections (around 10^5 times more frequently than among non-connections) [84].

Not only do liking mechanisms maintain relationships and express affiliation, they are also predictive of highly sensitive private traits and attributes such as intelligence, happiness, religious and political views, ethnicity, sexual orientation and use of addictive substances. [78].

In addition, users' interests (in particular aesthetic preferences) conveyed in social network sites are capable of identifying personal characteristics such as prestige, differentiation, theatrical persona and authenticity [85].

On the other hand, every trace left on social media - pictures, posts, videos and comments - reaches a large number of unacquainted observers: "*[...] the audience layer sits beyond the weak ties layer. It is made up of strangers [that] can play constructive roles when they are activated*" [119]. Many people "*use websites as a way to learn about someone they barely know*" [155]. Employers gathering information about job candidates are a typical case. According to the Harvard Business Review, the outcome of the interviews depends, to a significant extent, on the impression that the employers develop by watching the online material posted by the candidates [23].

These examples show that it is important to investigate the interplay between, on the one hand, the observable traces people leave online and, on the other hand, the impressions that these traces convey. For this reason, this chapter investigates the relationship between "*faves*" - the pictures that Flickr users tag as favourite - and personality impressions. In the next section, there is a brief review of the existing literature related to personality inference from social media data.

6.2 Related Works

This section surveys the main works in the literature that aim to address the inference of personality traits from social media data.

Social media platforms, such as the social networking application Facebook, the video sharing platform Youtube and the image hosting site Flickr have achieved broad popularity and have had prominent influence on our daily life since their creation. As self-presentation and self-disclosure are the main motivations behind the use of such social media platforms [73], numerous studies have been carried out to investigate the interplay between social media contents (profiles, posts, pictures and likes) and personality traits. Table 6.1 contains a synopsis of such works in terms of data, approaches and results. The personality traits include both the self-assessed personality traits that target the real personality of an individual and

are assessed by the individuals themselves [5, 15, 24, 103, 55, 134], and the attributed ones that focus on the personality of an individual perceived by others and hence are assigned by observers [24, 35, 43, 134]. Consequently, according to the definition by [157], the prediction of self-assessed and attributed traits corresponds to Automatic Personality Recognition (APR) and Automatic Personality Perception (APP), respectively.

The approaches proposed in [55] predicted each of the Big-Five personality traits (self-assessed) of 167 Facebook users by analysing the information which was revealed in their online profiles. The feature set included structural features such as density of user's egocentric network, those derived from personal information such as relationship status, religion and education history, and language features obtained using Linguistic Inquiry and Word Count (LIWC) which categorises words through linguistic and psychological processes [139]. The regression approaches are based on M5 algorithms and Gaussian processes with a mean absolute error within 11%. However, in their Gaussian processes approach, no ARD kernel was used to learn automatically the influence of each feature on the personality traits. Instead, a separate correlation analysis was performed to analyse the relationship between features and personality traits. In a similar fashion, the task of [5] is the discrimination between individuals scoring in the lower, middle or upper third of observed trait score ranges for all Big-Five traits. The subjects for study were 209 users of RenRen, a popular social networking platform in China. The features adopted included basic information (e.g. gender and age), usage statistics (e.g. blog usage frequency, data upload frequency), time-related features (e.g. status, count of published blog per recent one month) and emotion-related features (e.g. count of angry state within one blog). In this work, the C4.5 Decision Tree [117] achieved a classification accuracy of up to 72% (in terms of F-measure) depending on the trait.

The works of [15, 103] attempted to recognise personality traits without the expensive and time-consuming collection of self-assessments. This might be useful in scenarios where the population of users is large and hence collecting questionnaires might be difficult. [15] analyses posts from 156 users on FriendFeed, an Italian social network. The features are based on linguistic factors regarding topic (e.g. posts about job and music), word usage (e.g. the use of negative particles and word count) and psychological aspects (e.g. positive or negative emotions). The proposed unsupervised methodology measures first how stable the features

Table 6.1 APR and APP on social media.The table reports, from left to right, the number of subjects involved in the experiments, number and type of behavioural samples, main cues, type of task and performance over different traits. LIWC stands for Linguistic Inquiry Word Count (a psychologically oriented text analysis approach). The column "Other" refers to works using models different from the Big-Five. R stands for regression, U stands for unsupervised classification, C(n) for classification with n classes, and CA for correlation analysis. The performance is expressed in terms of Mean Absolute Error (MAE), F-Measure (F), accuracy (ACC) and correlation (ρ). As the results have been obtained over different data, the performances are not reported for comparison purposes, but to provide full information about the works described.

Reference	Subj.	Samples	Features	Task	Ext.	Agr.	Con.	Neu.	Ope.	other
[55]	167	167 Facebook Profiles	Egocentric networks, profile information, LIWC	R	0.12 MAE	0.10 MAE	0.10 MAE	0.11 MAE	0.10 MAE	
[5]	209	209 RenRen profiles	profile information, usage statistics, emotional states	C(2) C(3)	83.8 71.7 F	69.7 72.3 F	82.4 70.1 F	74.9 71.0 F	81.1 69.5 F	
[15]	156	473 posts on FriendFeed	some LIWC categories	U						average ACC 63.1
[103]	10000	10000 blog posts	LIWC	C(2)	80.0 ACC					
[43]	440	440 pictures	photo content, appearance	CA						see text
[35]	5216	5216 social media profiles	presence of personal information	CA						see text
[24]	300	60000 favourite pictures	visual pattern, aesthetic preferences	R	0.19 ρ	0.17 ρ	0.22 ρ	0.12 ρ	0.17 ρ	
[134]	300	60000 favourite pictures	visual characteristics	R						see text

are across multiple posts of the same user and then assigns personality traits according to the most stable features (average accuracy 63.1%). Similarly, the work of [103] simply labels users of blogging site Livejournal as extroverts or introverts based on the number of friends they have (extroverts have 108 to 150 connections whereas introverts have only 1 to 3 connections). Using the LIWC features, the prediction model involves two separate approaches. An SVM classifier is employed to predict the personality traits while logistic regression is used to investigate the predictive effect of each feature.

Compared to the earlier works aimed at addressing the APR problem, the methodology proposed in [43, 35] focuses on APP. In both works, features are extracted from profile information of users of social-networking websites and correlation analysis is utilised to explore the rater-target impression agreement, i.e the agreement between attributed traits by raters and self-assessed traits by targets (users). In particular, [43] analysed profile photographs and the results showed that pictures where the profile owners were smiling and outdoors correspond to the higher agreement. By contrast, [35] targeted personal information in the profiles and the experiments showed that the agreement is higher when users state their spirituality, beliefs, embarrassing and proud moments and when they post funny material.

The approaches proposed in [24, 134] address both APR and APP using the PsychoFlickr Corpus - the same data set used in the experiments of this work (see Section 6.3.1). Both works utilise variants of multiple instance regression [123] with Lasso [141] for prediction and an individual correlation analysis to capture the covariation between features and traits.

As the above survey shows, most of the works for personality inference focus on non-probabilistic parametric (various regression models) or non-parametric models such as SVM and C4.5 Decision Tree that give no indication about the predictive effect of each feature. Some works (e.g. [103]) had to combine non-parametric (SVM) and parametric (logistic regression) methods to make predictions and gain some insight into the influence of features on prediction. The experiments in this work attempted to exploit non-parametric Gaussian processes models to map favourite pictures into personality traits. This was possibly the first work to employ a non-parametric and fully probabilistic approach to infer personality traits from online pictures and learn automatically the interplay between features and personality factors during the training process.

6.3 Data and Personality

6.3.1 Data

The experiments in this work have been performed over *PsychoFlickr*, a publicly available corpus of $60,000$ pictures tagged as favourite by 300 Flickr users (200 faves per user), the subjects hereafter [24, 134]. For every subject, the Corpus includes two personality assessments (see Section 6.3.2): the first is the average of the traits attributed by 11 British assessors, the second is the average of the traits attributed by 11 Asian assessors. This makes it possible to investigate cultural effects.

6.3.2 Personality and Its Assessment

Personality is the latent construct that accounts for "*individuals' characteristic patterns of thought, emotion, and behaviour together with the psychological mechanisms - hidden or not - behind those patterns*" [47]. A large number of personality models have been proposed in the literature (an extensive survey can be found in [94]) and in this work, the personality assessments are presented in terms of the *Big Five Traits* (BF). The BFs are as follows: *Openness* (tendency to have wide interests and to be intellectually curious), *Conscientiousness* (tendency to be responsible, thorough and planful), *Extraversion* (tendency to be active and establish social relationships), *Agreeableness* (tendency to act according to the benefit of others) and *Neuroticism* (tendency to experience only the negative side of life).

The rationale behind the choice of the BF model is that, according to the literature on psychology, these are five behavioural dimensions that are known to capture most individual differences and are recognised as the most effective personality model proposed so far [132]. Consequently, the BF model is ubiquitous in personality computing [157] as well as personality science [165].

Moreover, from a computing point of view, the main advantage of the BF is that it represents personalities as five-dimensional vectors, a format particularly suitable for computer processing. Each component of the vector is a score that accounts for how well the behaviour of

an individual fits the tendencies associated with a particular trait. The scores can be obtained with questionnaires designed for personality assessment.

In these experiments, the questionnaire used was the *Big Five Inventory 10* (BFI-10), shown in Table 6.2. It is a psychometric instrument aimed at assessing individuals along the BF, and is abbreviated from the Big Five Inventory (BFI-44) to include only ten items of the original questionnaire [120]. The selected ten items have substantial correlations with the measurements obtained using the full forty-four items. This contributes to the major advantage of BFI-10: it takes less than one minute to fill in the questionnaire while still retaining significant levels of reliability and validity. Furthermore, a self-assessment questionnaire can be easily obtained by simply replacing "This person" with "I" in Table 6.2.

The five-point Likert scales (from "*Strongly Disagree*" to "*Strongly Agree*") are associated to each of the ten items, and mapped into the numbers ranging from -2 to 2. For a particular trait, the perceived personality scores can be computed from the assessors' answers to the corresponding items as below where Ai denotes the answer to question i:

- Extraversion: A1 (reversed-scored), A6
- Agreeableness:A2, A7 (reversed-scored)
- Conscientiousness: A3 (reversed-scored), A8
- Neuroticism: A4 (reversed-scored), A9
- Openness: A5 (reversed-scored), A10

where "reversed-scored" means that the numerical scoring scale runs in the opposite direction for the negatively worded questions.

Thus the integer score of each trait is in the interval [-4, 4]. Once the BF scores are available for all persons in a corpus, it is possible to estimate the median for each trait. In this way, the subjects can be split into two classes; namely those who are above median and the others. Hereafter, the classes are referred to as *high* and *low*, respectively.

Table 6.2 The BFI-10 questionnaire [120].

No.	Question	Disagree strongly	Disagree a little	Neither agree or disagree	Agree a little	Agree strongly
1	This person is reserved	(1)	(2)	(3)	(4)	(5)
2	This person is generally trusting	(1)	(2)	(3)	(4)	(5)
3	This person tends to be lazy	(1)	(2)	(3)	(4)	(5)
4	This person is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
5	This person has few artistic interests	(1)	(2)	(3)	(4)	(5)
6	This person is outgoing, sociable	(1)	(2)	(3)	(4)	(5)
7	This person tends to find fault with others	(1)	(2)	(3)	(4)	(5)
8	This person does a thorough job	(1)	(2)	(3)	(4)	(5)
9	This person gets nervous easily	(1)	(2)	(3)	(4)	(5)
10	This person has an active imagination	(1)	(2)	(3)	(4)	(5)

Table 6.3 Synopsis of the features.

Category	Name	*d*	Short Description
G1	Faces	1	Number of faces
G2: Colour properties	HSV statistics	5	Average of S channel and standard deviation of S, V channels [89]; *circular variance* in HSV colour space [91]; *use of light* as the average pixel intensity of V channel [27]
	Emotion-based	3	Measurement of *valence, arousal, dominance* of the emotions evoked by the colours [89, 152]
	Variety of colours	1	Colour diversity (colourfulness) measure based on Earth Mover's Distance (EMD) [27, 89]
G3	Colour distribution	11	Fraction of pixels that can be mapped into each of the 11 basic colour categories (*red, yellow, pink ,black, blue, brown, green, gray, orange, purple, white*) [89]
G4: Homogeneous regions	Edge pixels	1	Fraction of edge pixels in an image
	Level of detail	1	Number of homogeneous regions (after mean shift segmentation) [52, 21]
	Average region size	1	Average size of the homogeneous regions (after mean shift segmentation) [52]
	Image size	1	Size of the image [27, 88, 18]
G5: Composition	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [27, 89]
	Rule of thirds	2	Average of S,V channels over inner rectangle [27, 89]
G6	Texture Wavelets	12	Wavelets coefficients: Level of spatial graininess measured with a three-level Daubechies wavelet transform on HSV channels [27]
G7	GIST filters	24	Output of GIST filters for scene recognition [107].
G8	Gray Level Co-occurrence Matrix (GLCM)	12	Statistics of pixel values co-occurrences in 3×3 patches: amount of *contrast, correlation, energy, homogeneousness* for each HSV channel [89]
G9: Texture statistics	Tamura features	3	Amount of *coarseness, contrast, directionality* [138]
	Gray distribution entropy	1	Image entropy [88]

d denotes dimension, i.e. the number of features.

6.4 Feature Extraction

Deep convolutional neural networks (DCNN) [133] have been widely applied to learn features automatically from pictures. However, features learned by DCNN are generally difficult to interpret. As discussed in Chapter 1, being able to identify the relative influence of each feature on the prediction outcome, is of great importance for social signal processing applications. Therefore, this section presents a full description of the extraction of low-level features from online Flickr pictures, a synopsis of which can be found in Table 6.3.

Every picture of the corpus was represented by a set of 82 features inspired by Computational Aesthetics, the domain aimed at predicting whether people consider an image visually appealing or not [65]. The main reason behind this choice was that these features captured the visual appearance of the faves and this was the only information that the assessors had to attribute personality traits to the 300 subjects of the Corpus. Furthermore, the features have been shown to be effective in tasks similar to the one addressed in this work [24, 134]. In view of the G-ARD approach (see Section 6.5.1), the features (82 in total) have been split into 9 groups corresponding to the main visual properties of a picture (see the synopsis of the features in Table 6.3).

The feature set is designed to account for content independent visual characteristics and, hence, cope with the wide semantic variability of the pictures posted online. The only content dependent feature is the number of human faces (Group G1) because these are ubiquitous in pictures and furthermore, certain neural pathways make the human brain especially sensitive to face detection [71]. A subject of the PsychoFlickr Corpus is represented with the average of the 200 feature vectors extracted individually from every fave. In this way, the whole PsychoFlickr Corpus is represented by 300 vectors - one per subject (see Section 6.3.1). A full introduction of the group features in Table 6.3 is presented as follows:

G1: Number of faces

The number of faces, the only feature that takes into account the content of the images (i.e. what the images show), was calculated manually from each of the $60,000$ faves. Every

visible face was counted without considering its size, scale, pose and the facial expressions it contained. Since a pilot analysis shows that the accuracy of the Viola-Jones face detector [160] is only 70% and this introduces noises to the model, automatic face detectors were not adopted.

G2: Colour properties

This section describes the group features relating to the colour properties represented by the HSV colour space. HSV is an acronym for Hue, Saturation and Value (also referred to as Brightness).

HSV statistics: These features are obtained by collecting statistics over H, S and V pixel values of a picture, and they account for the use of colours. Let I, J denote the height and width of the image, respectively.

circular variance R [91] provides information of colour diversity and is computed as:

$$X = \sum_{i=1}^{I}\sum_{j=1}^{J} \cos H_{ij}, \quad Y = \sum_{i=1}^{I}\sum_{j=1}^{J} \sin H_{ij}$$
$$R = 1 - \frac{1}{IJ}\sqrt{X^2 + Y^2}$$

where H_{ij} is the Hue of pixel (i, j).

As light exposure discriminates well between aesthetically appealing and unappealing images while saturation indicates chromatic purity [27], the features corresponding to such vital observations of image aesthetics are computed as the average pixel intensity across the V and S channels:

$$\bar{V} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} I_V(i,j), \quad \bar{S} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} I_S(i,j) \tag{6.1}$$

where $\bar{V}$ is called *use of light*, and $I_V(i,j), I_S(i,j)$ denote the intensity value of pixel (i,j) across the V, S channel respectively.

Standard deviation is also calculated on the S and V channels. The average of the Hue was not computed because the interpretation of such a feature is not clear, because Hue is defined in the HSV space in terms of angles in a colour wheel.

Examples of how the figures change with the HSV statistics are shown in Figure 6.2.

Emotion-based: According to psychological studies, Saturation and Brightness show evidence of strong and consistent effects on emotions [152], and emotional reactions to them are usually expressed through the Pleasure-Arousal-Dominance emotion model (where pleasure is referred to as valence in this work):

$$\text{Valence} = 0.69\bar{V} + 0.22\bar{S}$$
$$\text{Arousal} = -0.31\bar{V} + 0.60\bar{S}$$
$$\text{Dominance} = -0.76\bar{V} + 0.32\bar{S}$$

where $\bar{V}, \bar{S}$ are defined in eq. (6.1). Figure 6.2 provides examples of pictures with different levels of Valence, Arousal and Dominance.

Variety of colours: This feature accounts for relative colour distribution and distinguishes multi-coloured images from mono-chromatic, sepia or simply low contrast images. After conversion to the **CEILUV** colour space, the Earth Mover's Distance (EMD) [129] is used to measure the similarity between the image under analysis and an ideal colourful image according to the algorithm proposed in [27, 89].

G3: Colour distribution

Every pixel in the image is mapped onto one of the eleven basic colour terms (black, blue, brown, gray, green, orange, pink, purple,red, white and yellow) [8] using the algorithm of [162]. Then the fraction of pixels assigned to each of the colour terms is calculated as the feature. This feature, on the one hand, can account for the frequency of occurrences of the colours in the image and, on the other hand, can capture the style of a photographer.

G4: Homogeneous regions

Edge pixels: Pixels of sharp changes in brightness (called edge pixels) are often interesting as they indicate object boundaries and other kinds of meaningful changes such as reflectance changes (e.g. stripes of zebras and spots of leopards) and illumination changes (e.g. cast

shadows) [45]. Consequently, the Canny detector [88] is adopted to identify the edges and the fraction of edge pixels in an image is computed as a feature. The top-left of Figure 6.3 (with title "Canny") provides an example of edge extraction in an image.

Objects and scene semantics have been shown to play a very important role in understanding the subjective judgement of a picture [26, 70]. Following this, image segmentation was performed to collect low-level statistics. The EDISON implementation [52] of the mean shift segmentation algorithm [21] was employed. After segmentation of an image, the number of segments is calculated as the **Level of detail** feature, accounting for the regions "density" of an image. The normalised **Average region size** of the homogeneous regions has also been collected as one feature. Normalisation is achieved by dividing the mean size of the regions by the size of the whole image. As is shown in the top-right of Figure 6.3 (with title "Level of detail"), a higher number of segments provides more details.

Image size: This feature accounts for the total number of pixels in an image.

G5: Composition

Composition analysis is concerned with analysing the spatial relations between the visual elements of a picture [89]. In this work, the following two aspects of composition are considered.

Low depth of field (DOF) is often used by professional photographers to blur the background and make the object of interest noticeably sharper in order to draw the attention of the observer [27, 89] (see bottom-left of Figure 6.3). Following [27], the image is divided into 16 equal rectangular blocks $\{M_1, ...M_{16}\}$ numbered in row-major order. Let $w_3 = \{w_3^{hl}, w_3^{lh}, w_3^{hh}\}$ denote the set of high-frequency (Level 3 by the notation used in eq. (6.4)) wavelet coefficients (see the following section) of the hue image I_H. The reason for choosing Level 3 is that, in an image with low DOF, the object of interest is assumed to be central and high-frequency co-efficients encode fine visual details. The feature that indicates the low depth of field is computed as follow:

$$DOF_H = \frac{\sum_{(i,j)\in M_6 \cup M_7 \cup M_{10} \cup M_{11}} w_3(i,j)}{\sum_{k=1}^{16} \sum_{(i,j)\in M_k} w_3(i,j)} \tag{6.2}$$

i.e the ratio of the wavelet coefficients in the high-frequency of the inner part of the image to those of the whole image. Similarly, this low depth of field indicator is computed for the S, V channels of the image. The lower part of Figure 6.3 with the title "Low depth of field indicator" shows how the image changes with different DOF.

Rule of thirds is a very popular photography composition guideline. It suggests that the main object (centre of interest) in a photograph should be positioned at one of the four intersections or along one of the lines on the inside as shown in Figure 6.1. The 'rule of thirds' feature is obtained by computing the average values of Brightness and Saturation of the inner rectangle [27, 89]:

$$f_V = \frac{9}{IJ} \sum_{i=I/3}^{2I/3} \sum_{j=J/3}^{2J/3} V_{ij} \tag{6.3}$$

where I is the image height, J is the image width and V_{ij} is the Brightness at pixel (i, j). A similar feature f_S can be computed for the Saturation.

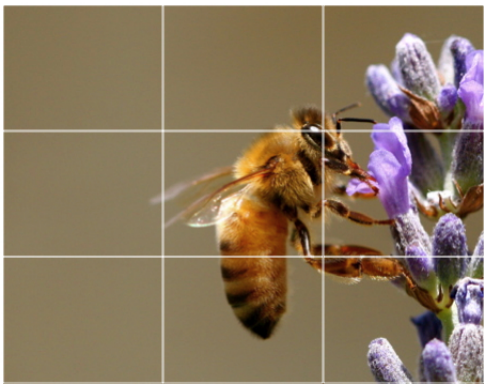


Figure 6.1 The rule of thirds guideline in photography: an image is ideally divided horizontally and vertically each into three parts. Important parts of the composition are placed at the intersection points instead of the centre. http://digital-photography-school.com/rule-of-thirds/.

G6: Texture wavelets

Texture is defined as "*the set of local neighbourhood properties of the grey levels of an image region*" [86]. It accounts for the intuitive properties of images such as roughness, granulation and regularity and thus serves as a significant cue for images analysis. Daubechies wavelet

transform [28] can be employed to measure spatial smoothness/granulation in the image [27, 89].

In this work, a *three-level* 2D Discrete Wavelet Transform (2D-DWT) was performed on all three H, S, V channels, where high frequency was associated to high edge density. Figure 6.5 provides an example of a *two-level* wavelet transform. As is shown in Figure 6.5(b), the *two-level* wavelet bands have been arranged from upper left to lower right in the transformed image. At each level, the four bands were labelled by LL (LowLow), HL(HighLow), LH (LowHigh), HH(HighHigh) (see Figure 6.5(a) for their arrangement). LL part is a low-pass version of the original image, whereas HL, LH, HH parts correspond to images with their horizontal, vertical and diagonal edges at the finest scale highlighted, respectively. Let $w_k^{hl}, w_k^{lh}, w_k^{hh}$ denote the wavelet coefficients at level k ($k = \{1, 2, 3\}$) for the H channel. The corresponding wavelet feature is derived from the following equation:

$$wf_k = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} w_k^{hl}(i,j) + \sum_{i=1}^{I}\sum_{j=1}^{J} w_k^{lh}(i,j) + \sum_{i=1}^{I}\sum_{j=1}^{J} w_k^{hh}(i,j)}{(\left|w_k^{hl}\right| + \left|w_k^{lh}\right| + \left|w_k^{hh}\right|)} \tag{6.4}$$

where wf_k denotes the wavelet feature at level k. Wavelet features for the S, V colour space channels are computed similarly. Therefore, a *three-level* wavelet transform will result in nine features (three features for each level). At each level, the sum of wf_k for the three channels has also been calculated, leading to another three features.

G7: GIST filters

It is a low dimensional representation of the structure of real world scenes termed as *Spatial Envelope*. The spatial envelope relies on Gabor Filters to capture a set of perceptual properties, namely naturalness, openness, roughness, ruggedness and expansion [107]. The outputs of the GIST filters are used as features.

G8: Gray Level Co-occurrence Matrix (GLCM)

The GLCM is a matrix where the entry (m, n) is the probability $p(m, n)$ of observing values m and n for a given channel (H, S or V) among the pixels in the same region R. In the process of feature extraction, R includes a pixel and its right neighbour and, consequently,

the GLCM includes the probabilities of observing one pixel with the value n that is at the right of a pixel with the value m. Several features have been calculated based on GLCM, each of which has been obtained individually over the H, S and V channels [61]:

Contrast is computed as the average value of $(m-n)^2$, the square difference of values observed in neighbouring pixels:

$$C = \sum_{m,n=0}^{L-1} (m-n)^2 p(m,n) \tag{6.5}$$

where L is the number of possible values of a pixel. The minimum of C is 0, corresponding to a uniform image, whereas the maximum of C is $(L-1)^2$. Figure 6.4 (the third row) provides an example of pictures with high and low contrast.

Correlation is the coefficient that accounts for the measurement of the covariation between neighbouring pixels:

$$\sum_{m,n=0}^{L-1} \frac{(m-\mu)(n-\mu)p(m,n)}{\sigma^2} \tag{6.6}$$

where $\mu = \sum_{m,n=0}^{L-1} mp(m,n)$, and $\sigma^2 = \sum_{m,n=0}^{L-1} p(m,n)(m-\mu)^2 + \sum_{m,n=0}^{L-1} p(m,n)(n-\mu)^2$. The correlation ranges between -1 and 1 (see the third row of Figure 6.4 for an example of pictures with different correlation).

Energy is computes as the sum of the square values of the GLCM entries:

$$\sum_{m,n=0}^{L-1} p(m,n)^2 \tag{6.7}$$

The energy of an uniform image is 1. Bottom-left of Figure 6.4 gives an example of pictures with high and low energy.

Homogeneity measures how frequently neighbouring pixels have the same value (see bottom-right of Figure 6.4 for an example) :

$$H = \sum_{m,n=0}^{L-1} \frac{p(m,n)}{1+|m-n|} \tag{6.8}$$

Larger elements on the diagonal of the GLCM tend to result in higher homogeneity.

G9: Texture statistics

Tamura features:

[138] proposed six texture features approximating human visual perception, namely, coarseness, contrast, directionality, line-likeness, regularity and roughness. Only the first three have been considered in this work, as they have been found to be tightly correlated with human perception.

Coarseness accounts for the size of texture elements (texels). The essence of this approach is to pick large size texels for coarse texture but to pick small size texels for fine texture. The following steps summarise this procedure:

1) Compute the average at every point over neighbourhoods whose size are powers of two, e.g. $1 \times 1, 2 \times 2, 4 \times 4, ..., 32 \times 32$. The average over the neighbourhood of size $2^k \times 2^k$ at point (i, j) takes the form:

$$A_k(i,j) = \sum_{m=i-2^{k-1}}^{i+2^{k-1}-1} \sum_{n=j-2^{k-1}}^{j+2^{k-1}-1} f(m,n)/2^{2k} \tag{6.9}$$

where $f(m, n)$ is the grey-level at (m, n).

2) For each point, compute the difference between non-overlapping neighbourhoods on opposite sides of the point in both horizontal and vertical directions:

$$E_k^h(i,j) = |A_k(i + 2^{k+1}, j) - A_k(i - 2^{k+1}, j)| \tag{6.10}$$

and

$$E_k^v(i,j) = |A_k(i, j + 2^{k+1}) - A_k(i, j - 2^{k+1})| \tag{6.11}$$

where $E_k^h(i, j), E_k^v(i, j)$ denote the difference for the horizontal and vertical orientation, respectively.

3) At each point, pick the size that gives the largest difference value:

$$S_{best}(i,j) = 2^k$$

where k maximize E in either orientation, i.e.

$$E_k = \max_{d=h,v}(E_1^d, E_2^d, ..., E_5^d)$$

with $k \in \{1,2,3,4,5\}$.

4) Finally take the average of $S_{best}(i,j)$ over the image as coarseness measure:

$$F_{crs} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} S_{best}(i,j) \tag{6.12}$$

where I and J are the width and height of the image.

For an example of pictures with different values of coarseness, see top-right of Figure 6.4.

Contrast stands, in the narrow sense, for picture quality. In practice, the following two factors influence the contrast: dynamic range of grey-levels (the larger the range, the higher the contrast), polarisation of the distribution of black and white on the grey-level histogram (pictures with high contrast have polarised histogram). It is computed as follows:

$$F_{con} = \frac{\sigma}{\alpha_4^z} \qquad with \qquad \alpha_4 = \frac{\mu_4}{\sigma_4} \tag{6.13}$$

where μ_4 is the fourth moment about the mean and σ is the standard deviation of grey-levels of the picture. z has experimentally been chosen to be $\frac{1}{4}$ to yield the best result. See the second row of Figure 6.4 for an example of pictures with high and low contrast.

Directionality measures how polarised the distribution of edge directions is. High directionality accounts for a texture with homogeneously oriented edges whereas low directionality indicates a texture where the edges are heterogeneously oriented. It has been calculated in the following way: First, calculate the entropy E of the distribution of the directions of all the edge pixels. Then directionality is measured as $1/(E+1)$. The distribution of textures with edges oriented along a single direction features sharp peaks, resulting in $E = 0$ and a

maximal directionality of 1. Conversely, images with a nearly flat distribution of edge orientations will have low directionality (≈ 0). See the second row of Figure 6.4 for an example of pictures with different levels of directionality.

Grey-level distribution entropy measures the homogeneousness of an image. The computation of this feature involves the following steps. First, convert the image into grey-levels. Then, for every pixel, calculate the distribution of the grey-values in a neighbourhood of 9×9 pixels, i.e. the grey-level histogram of the patch. After that, compute the entropy (a measure of the average amount of information) of the distribution. Finally, sum up all the entropy values and divide by the size of the image. A uniformly distributed intensity of an image will give a low entropy (see top-left of Figure 6.4 for an example of how the entropy impacts visual characteristics). The entropy takes the form:

$$E = -\sum_{i} P_i Log_2 P_i \tag{6.14}$$

where P_i is the probability that the difference between two adjacent pixels is equal to i, and Log_2 is the base 2 logarithm.

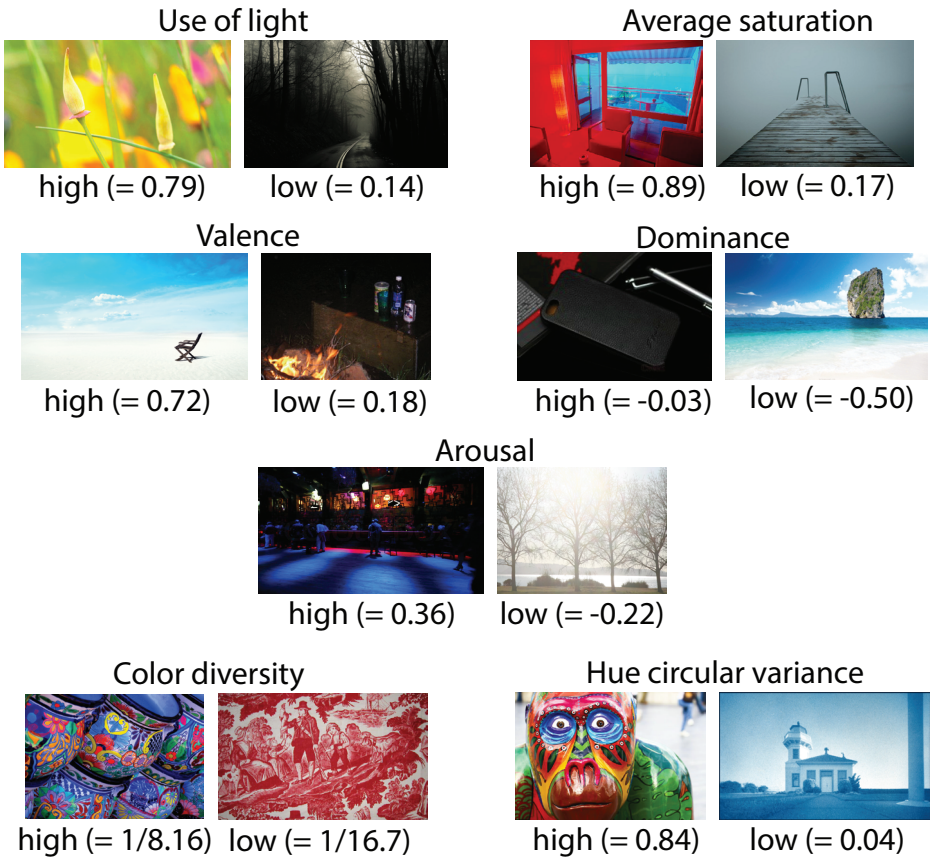


Figure 6.2 Examples of how visual properties of a picture change with several colour-related features.

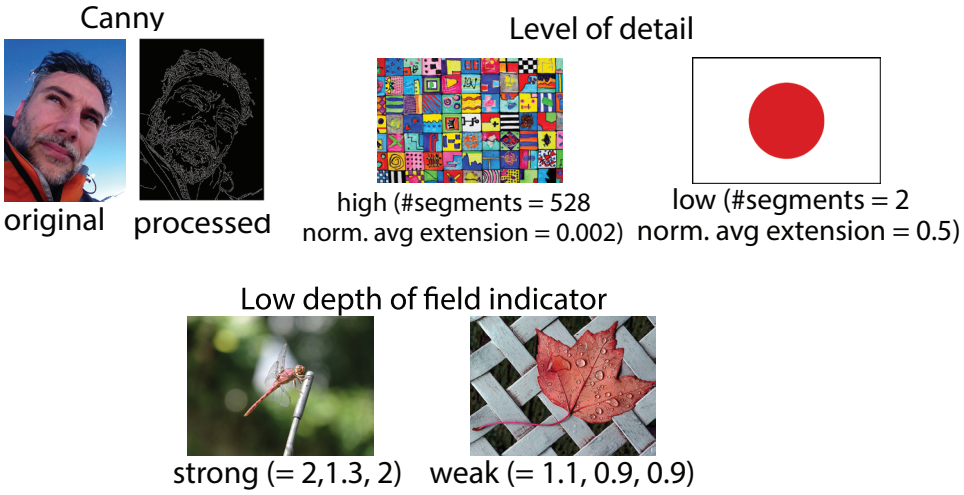


Figure 6.3 The upper-left panel of the figure shows the effect of the Canny algorithm and the rest shows the visual properties related to Level of Detail and Low Depth of Field.

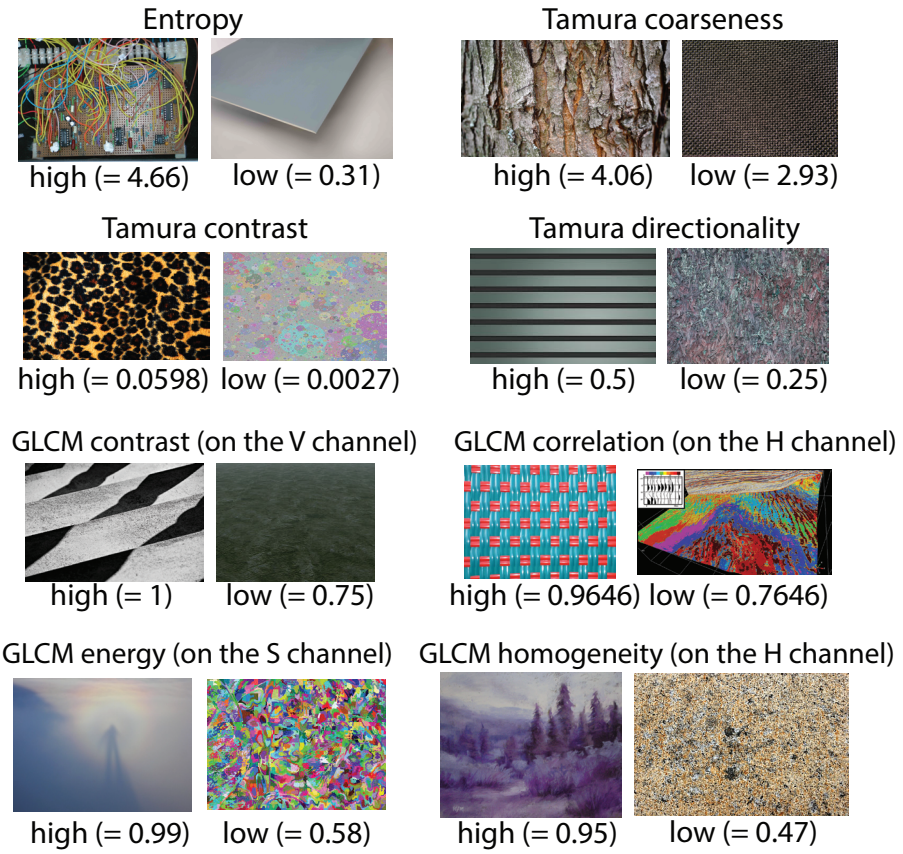


Figure 6.4 Examples of the textual properties associated to *G8, G9*.

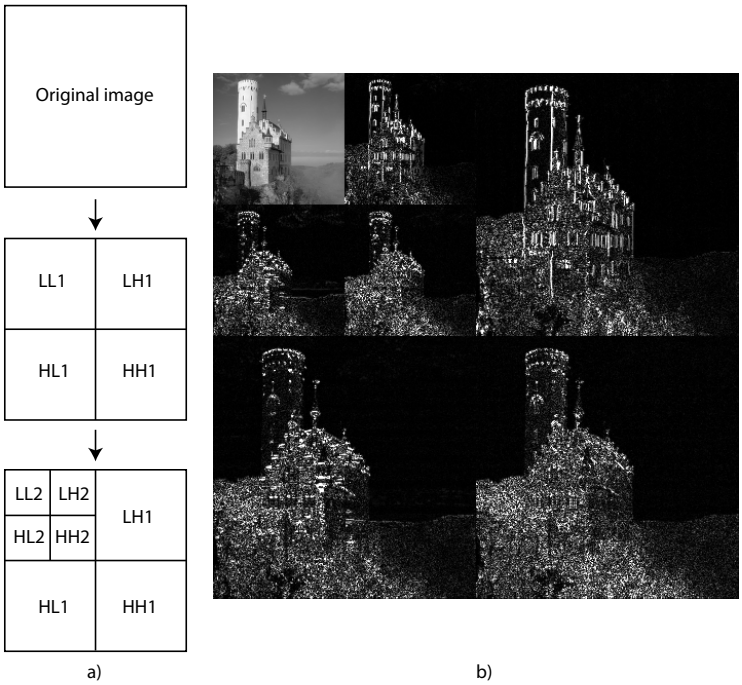


Figure 6.5 Wavelet decomposition.

.

6.5 Inference of Personality Traits

This section describes in detail the approaches to classification of personality traits into *high* or *low* (see Section 6.3.2).

6.5.1 Trait Classification

This work proposes the adoption of a flexible and highly descriptive classification model based on Gaussian processes (GPs) [121]. These share with support vector machines (SVM, see Section 6.5.3) - the classifier that achieves state-of-the-art results in most tasks addressed in the literature - the important property of being non-parametric. However, GPs have at least two major advantages. The first is that an appropriate definition of their kernel allows an explanation of the role played by the different feature groups in the classification (without explicit knowledge of the mapping between features and labels). The second is that GPs are formulated in probabilistic terms and, hence, a Bayesian treatment allows an incorporation of confidence levels when making predictions.

Under the GP assumption (see Chapter 2), the latent values $\mathbf{f}$ are jointly Gaussian distributed with $p(\mathbf{f} \mid \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K}$ is the kernel matrix and $\boldsymbol{\theta}$ is the parameter vector of the kernel function. The main novelty introduced in this chapter is the *Group-Automatic Relevance Determination* (G-ARD) kernel function, a new kernel parametrisation designed to quantify the role played by the feature groups identified in Section 6.4 or any other meaningful partitions of the features:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp\left\{ -\sum_{r=1}^{N_g} \frac{1}{N_r {\tau_r}^2} \left[\sum_{s \in \mathcal{G}_r} (\mathbf{x}_{i(s)} - \mathbf{x}_{j(s)})^2 \right] \right\}, \tag{6.15}$$

where σ is the marginal variance of the latent values, τ_r is the length-scale parameter for group r (it ensures, on the one hand, that the weights do not depend on the number of features in the groups and, on the other hand, that different weights are comparable even if the respective groups include different numbers of features), N_r is the number of features in group r, N_g is the number of groups, $\mathbf{x}_{i(s)}$ is the s^{th} component of vector $\mathbf{x}_i$, and $\mathcal{G}_r$ is the set of the indices of the features that belong to group r. This formulation of the G-ARD kernel

can be interpreted as product of multiple kernels imposed on the different feature groups of particular interest. Compared to the ARD kernel, the G-ARD allows the reduction of the number of weights from 82 (the number of features) to 9 (the number of groups).

6.5.2 Fully Bayesian Inference of Parameters and Predictions

Let $\boldsymbol{\theta}$ denote the parameter vector of the G-ARD kernel function. Given a new input vector $\mathbf{x}_\star$ with latent value $f_\star$, a fully Bayesian treatment of the $y_\star$ prediction requires a solution to the following integral:

$$p(y_\star \mid \mathbf{y}) = \int p(y_\star \mid f_\star) p(f_\star \mid \mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y}) df_\star d\mathbf{f} d\boldsymbol{\theta} \tag{6.16}$$

where $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$ is the posterior over $(\mathbf{f}, \boldsymbol{\theta})$. In contrast to a point estimate of $\boldsymbol{\theta}$ for prediction, which may potentially underestimate uncertainty or cause inaccurate evaluation of the relative influence of different features [75], the posterior $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$ encodes the uncertainty in model parameters and thus enables an understanding of the importance of different features with confidence on it. Since the computation of eq. (6.16) is analytically intractable, Monte Carlo approximation methods are usually employed.

Sampling from $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y})$ is highly non-trivial, and it is normally done by means of Markov Chain Monte Carlo (MCMC) algorithms [40]. In this work, predictions are carried out using an adaptive importance sampling-based approach, and in particular the pseudo-marginal adaptive multiple importance sampling (PM-AMIS) proposed in Section 4.3. The motivation is that such a methodology has been shown to be faster (see Section 5.5.2), compared to state-of-the-art MCMC approaches, in computing predictions for GP models. The intuition is that the algorithm adaptively constructs an approximate posterior over $(\mathbf{f}, \boldsymbol{\theta})$ that is used to build an increasingly more accurate importance sampling estimator of the predictive distribution above. The importance weights have the following form:

$$w_i^t = p(\boldsymbol{\theta}_i^t) / \frac{1}{\sum_{t=0}^{T-1} N_t} \sum_{t=0}^{T-1} N_t q_t(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}}_t), \tag{6.17}$$

where T is the total number of iterations, $p(.)$ denotes the posterior of $\boldsymbol{\theta}$ up to a constant, $q_t(.)$ denotes the importance density at iteration t with sequentially updated parameters $\widehat{\boldsymbol{\gamma}}_t$,

and $\boldsymbol{\theta}_i^t$ are samples drawn from $q_t(.)$ with $0 \leq t \leq T-1$, $1 \leq i \leq N_t$. At each iteration of PM-AMIS, all the importance weights get updated, including those computed at previous iterations. Because in GP classification the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ cannot be computed analytically, PM-AMIS resorts to an unbiased estimate of the marginal likelihood using a "nested" importance sampling estimation procedure. Even though the computation of the weights is now approximate, because $p(\mathbf{y} \mid \boldsymbol{\theta})$ is estimated unbiasedly, it can be shown that PM-AMIS does not introduce any bias in predictions (see Section 4.3).

6.5.3 Support Vector Machines

The Support Vector Machine (SVM) is one type of decision machine but does not provide posterior probabilities [10]. The name, SVM, comes from the fact that, if the data is non-degenerate, the separating hyperplane (decision boundary) is uniquely defined by $d+1$ support vectors where d is the dimensionality of the data. The distance between the decision boundary and the nearest data point (sample) is called the margin. This margin indicates confidence in the prediction (a sample belongs to a positive or negative class). The larger the margin, the more the confidence there is in the prediction. SVM chooses the decision boundary that maximises this margin, so it is often called a Large Margin Classifier. Consequently, in SVM, the decision boundary has the special property that it is as remote as possible from both the positive and the negative data points.

Let $\mathbf{x}_1, ..., \mathbf{x}_N$ be the input data points, $y_1, ..., y_N$ denote the corresponding labels where $y_n \in \{-1, 1\}$. A linear two-class discriminant function takes the form:

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{6.18}$$

where $\phi(\mathbf{x})$ denotes a fixed feature-space transformation, and b is the bias parameter (the negative of which is sometimes called a threshold). In case of a linearly separable training data set, we have $f(\mathbf{x}_n) > 0$ for the positive samples ($y_n = 1$) and $f(\mathbf{x}_n) < 0$ for the negative ones ($y_n = -1$). It can be shown that in this case the margin is defined by $\frac{1}{\|\mathbf{w}\|}$. Finding an

optimal value of $\mathbf{w}$ that maximises the margin $\frac{1}{\|\mathbf{w}\|}$ requires optimising the following:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \quad \textit{s.t.} \quad \forall i, y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 \tag{6.19}$$

Equation (6.20) is called SVM with "hard" constraints where *Lagrange multipliers* can be used to solve this kind of optimisation problem.

The above formulation makes the assumption that the data is linearly separable, i.e. it is always possible to find a hyperplane that will perfectly separate the positive and negative training samples. However, most of the real data sets are noisy, which means there is no such separating hyperplane. In most cases where the data cannot be nicely separated, a penalty (cost) has to be introduced, leading to the following formulation:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \mathbf{C}\sum_{i=1}^{n}\xi_i \quad \textit{s.t.} \quad \forall i, y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \tag{6.20}$$

where $\xi_i = \max\{0, 1 - y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b)\}$ is the empirical loss (also called the *Hinge Loss*) introduced when $\mathbf{x}_i$ is on the wrong side of the margin, the regularisation parameter $\mathbf{C}$ controls the loss (cost) for misclassification, i.e. it controls the trade-off between fitting the data (related to the "bias" property of a predictive model) and making the margin large (related to the "variance" property of a predictive model). In this way, apart from minimising $\frac{1}{2}\|\mathbf{w}\|^2$, the number of misclassifications (cost) is also minimised.

Kernels allow for complex, non-linear classifiers using Support Vector Machines by transforming the input feature vector $\mathbf{x}$ into a higher dimensional space. A typical Gaussian radial basis function kernel takes the following form:

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x}, \mathbf{x}'\|^2/2\sigma^2) \tag{6.21}$$

In this case, the regularisation parameter $\mathbf{C}$ and the Gaussian kernel parameter σ can be trained using cross-validation described in the following Section 6.5.4.

It should be noted that, although probabilistic outputs can be formulated for SVM [115, 54], the objective function therein [115] is somewhat artificial because it uses logistic regression to obtain a likelihood. In contrast, GPs offer an objective function by default. In addition,

despite optimising kernel parameters in SVM can also obtain information on the relative importance of the features, however, this is generally difficult or expensive to do [81, 46], whereas in GPs the kernel parameters are learned automatically during the training process.

6.5.4 Experimental Setup

In the experiments (see Section 6.6), the $k-$fold cross-validation method [77] is used to measure the predictive performance (expressed in terms of prediction accuracy) for both PM-AMIS and SVM described in Sections 6.5.2 and 6.5.3 respectively. The $k-$fold cross-validation involves randomly partitioning the entire data set into k non-overlapping groups (data "folds") of approximately equal size. Taking $k = 10$ for example, the original data D is split into 10 equal size subsets that are mutually exclusive: $D_1, ..., D10$, with one $D_i (i \in \{1, ..., 10\})$ designated for testing and the remaining nine folds for training. The overall accuracy has been computed as the average of the accuracies obtained over the ten partitions. In this way, the $k-$fold cross-validation corrects for the optimism of training error by averaging measures of fit (prediction error) and hence derives a more accurate estimate of prediction accuracy [57].

Table 6.4 Confusion matrix.

	Positive (Actual)	Negative (Actual)
Positive (Predicted)	True Positive (TP)	False Positive (FP)
Negative (Predicted)	False Negative (FN)	True Negative (TN)

The prediction accuracy in this work is measured in terms of the F-score [124], which conveys the balance between the precision (a measure of a classifier's exactness, also called *Positive Predictive Value*) and the recall (a measure of a classifier's completeness, also known as *Sensitivity*). Given the confusion matrix (also called a contingency table), shown in Table 6.4, several common performance metrics can be calculated as below [38]:

- recall $= \frac{\text{Positives correctly classified}}{\text{Total positives}} = \frac{\text{TP}}{\text{TP + FN}}$
- precision $= \frac{\text{TP}}{\text{TP + FP}}$
- F-score $= \frac{2}{\text{1/precision + 1/recall}}$

In this work, precision and recall are given equal weight resulting in the above formulation of F-score [83].

6.6 Results

The experiments addressed the task of predicting whether a subject belongs to class *high* or *low* for the Big-Five (see Section 6.3.2). The main reason behind this choice is that it corresponded to the natural tendency to compare others in terms of who is higher or lower along a given dimension: "*a compelling argument can be made for emphasising comparisons among individuals, which we do in everyday life (Who is more assertive? Who is more responsible?) and which is useful for such practical purposes as deciding whom to hire for a particular job*" [19].

The experiments were performed using a k-fold ($k = 10$) validation approach (see Section 6.5.4) and the same subject never appears in training and test set. The classification has been performed with both the G-ARD approach (see section 6.5.1) and an SVM with radial basis function kernel (one parameter, see Section 6.5.3). The SVM classifier optimised the kernel parameters by minimising the cross-validation error across a set of candidate values, which is, generally, not feasible for large parameter sets and / or small amounts of data such as the one adopted in this work (300 subjects in total). In contrast, the proposed G-ARD GPs can integrate the uncertainty in the kernel parameters by means of a Bayesian approach when they make predictions. As a result, Figure 6.6 shows that the G-ARD is competitive with the SVM even if the number of its kernel parameters is larger. The accuracy differences across the traits are in line with results typically observed in Personality Computing [157], where different traits can be predicted with different degrees of accuracy depending on the particular data. This parallels the psychological concept of *relevance* according to which traits emerge with different evidence in different contexts (e.g. extraversion is easier to observe at a party than at a funeral) [165].

Besides achieving high accuracy, the G-ARD provides information about the feature groups that better account for the classification outcomes. Figure 6.7 shows the inverse of the G-ARD weights ($1/\tau_r$ defined in eq. (6.15)) for both Asian and British assessors. Overall,

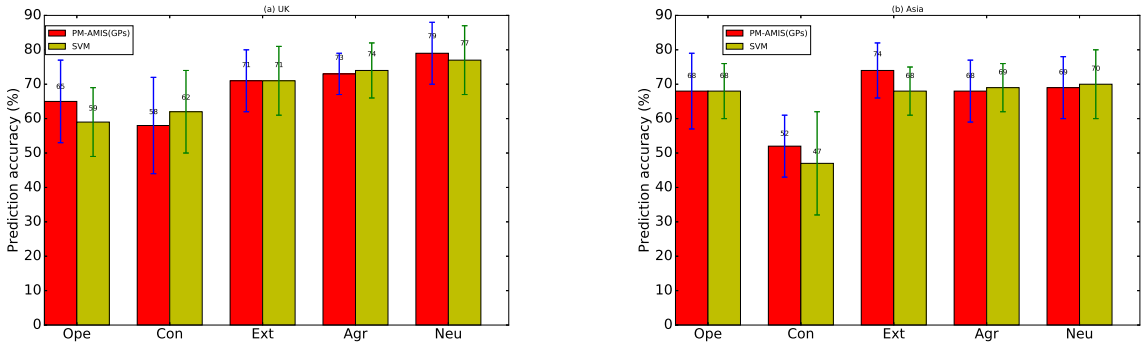


Figure 6.6 Prediction accuracy of PM-AMIS/SVM for the two cultures (UK and Asia).

the presence of human faces (group G1) plays the most important role for all traits and both cultures. The only exception is Extraversion, where the role of G1 is significant, but comparable to those of other groups. The probable reason is that in the case of this trait, strongly associated to social interactions, it is important not only that there are other faces, but also in what type of image they appear (e.g. the face is the main element in a portrait, but it is just a detail in the picture of a crowded public space). Overall, faces appear to be more important for British assessors than for Asian assessors for all traits except Neuroticism (the difference between the G1 weights is always statistically significant). The other feature groups for which the weights are large are those that correspond to high level aspects of a picture, namely amount and size of visually homogeneous regions (G4), composition (G5) and textural properties (G9). The other groups have a non-negligible role but appear to be less important. One possible reason of these results is that visual features accessible at first glance, such as those included in the groups above, are probably more likely than others to drive the personality impressions of the assessors.

The difference between the weights resulting from British and Asian assessors is always statistically significant except in the case of G5 for Neuroticism ($p < 0.01$ after Bonferroni Correction according to a weighted two-sample t-test). These results suggest that there is a cultural effect on personality perception. The largest differences can be observed for G1 (see above). Furthermore, British assessors are less sensitive to number and size of visually homogeneous regions for Agreeableness and Neuroticism while they are more sensitive to the texture properties for Conscientiousness and Agreeableness (conversely for Openness). Overall, Figure 6.7 suggests that UK and Asian assessors are sensitive to the same visual

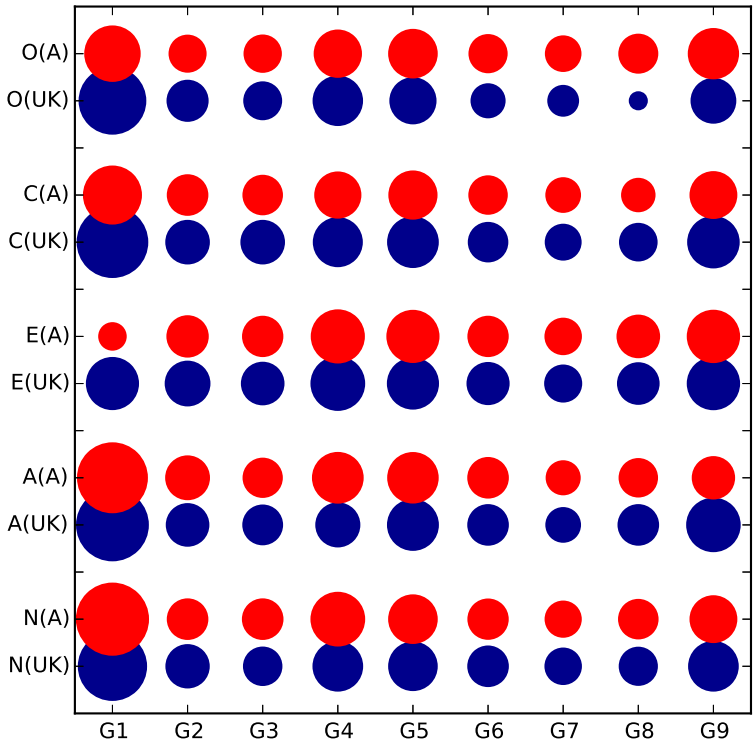


Figure 6.7 The plot shows the co-efficients of the G-ARD for the five traits (O,C,E,A,N) and the two cultures, namely Asia (A) and UK.

characteristics, but with different relative importance. One possible explanation is that there is no cultural difference for the physiological aspects - hence all assessors are sensitive to the same visual features - but there are cultural differences when it comes to the association between visual features and personality traits.

Another appealing property of the GPs compared to SVM is their capability of quantifying uncertainty resulting from their probabilistic formulation. The above Figure 6.7 shows the mean of the coefficients of the nine groups of features, where those of G3, G4, G5 and G6 are very similar for the British and Asian assessors. However, the t-test shows that the difference between the weights resulting from the British and Asian assessors are statistically significant. This difference is more obvious by looking at the density plot of the weights. Figure 6.8 shows the density plot of the co-efficient of G5 for the trait of Agreeableness (Agr) for both the British and Asian assessors. As can be seen, there is remarkable difference between the distributions of the co-efficient of G5 for the British and Asian assessors, reflecting the cultural effects between personality assessors of different national origin when associating the visual features with the personality traits.

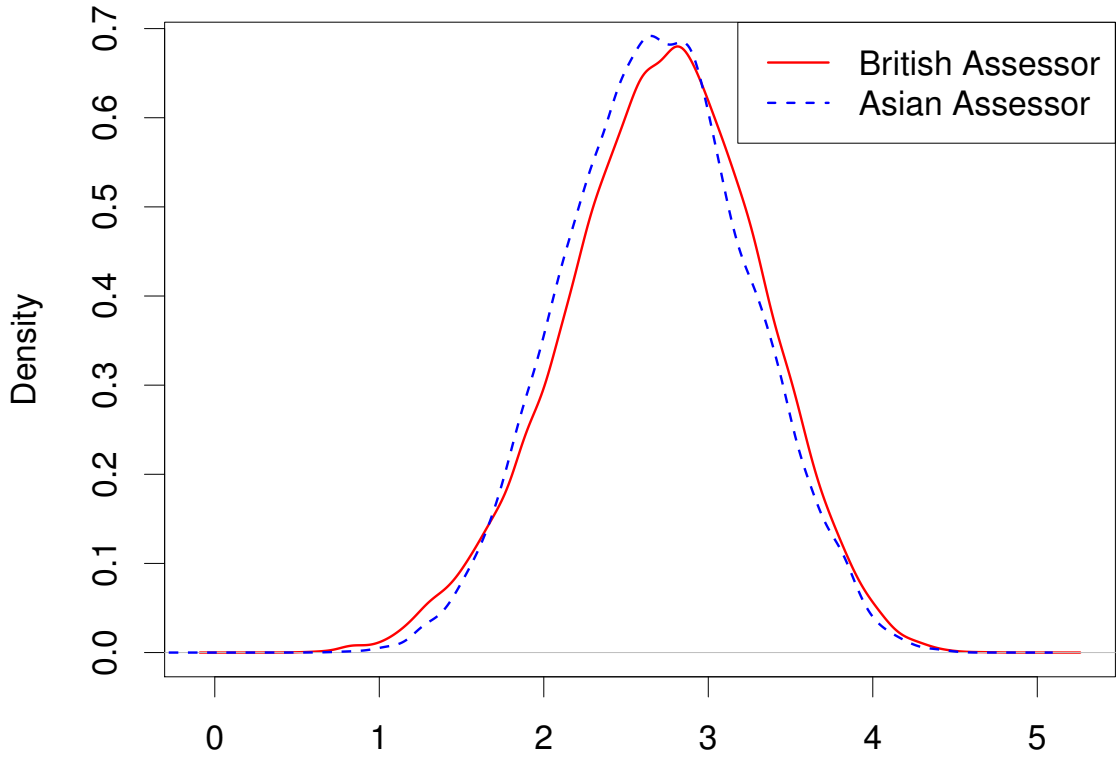


Figure 6.8 Density plot of the co-efficient of G5 for the trait of Agreeableness (Agr) for both the British and Asian assessors.

6.7 Conclusion

This chapter has shown that Flickr faves can be used to predict whether a Flickr user is perceived to be above the median with respect to the Big-Five traits. The results showed that the new G-ARD kernel designed for the experiments of this work allows a GP-based classifier to achieve comparable accuracies as state-of-the-art SVMs. Furthermore, the parameters of the G-ARD kernel allowed the identification of the groups of features that better account for the classification outcome while detecting cultural differences between UK and Asian personality assessors.

The classification accuracies, well above chance for all traits, showed that the weights of the G-ARD kernel provide reliable information about the interplay between low-level, visual characteristics of faves and attribution of personality traits. According to recent sociological investigations [119], this is important because the impression people convey online can

change the outcome of important issues like, e.g. getting or not getting a job [23]. For this reason, future work will concentrate on how to use the information provided by the G-ARD weights to ensure that items posted online do not convey a wrong impression, whether it comes to faves or other types of online material.

This chapter has demonstrated that the interplay between the interpretable features extracted and the prediction outcome can be learned through the G-ARD kernel, which is important as it contributes to the understanding of human behaviours. In the following chapter, the extraction of features using Hilbert spectral analysis (HSA) [131] will be explored. The motivations behind this choice of feature extraction method are as follows. On the one hand, HSA is a new representation of signal fundamentally different from STFT that dominates feature extraction in SSP. On the other hand, compared to features learned by DCNN that are generally hard to interpret, the resulting features are interpretable, which is of great importance for SSP.

Chapter 7

Feature Extraction Using Hilbert Spectral Analysis

In signal processing, the conventional way to complex extend a real signal in order to obtain the instantaneous amplitude (IA) and instantaneous frequency (IF) is to use the Hilbert transform to construct Gabor's analytic signal. This approach relies on the assumption of harmonic correspondence (HC), which may lead to incorrect IA and IF. Hilbert spectral analysis (HSA) [131] , a multi-component model that features a relaxation of the HC condition, is reported to be capable of generating exact IA and IF when appropriate assumptions are made on the signal model and thus provides a new and powerful framework for signal analysis. This chapter aims to explore whether features extracted using HSA can improve the accuracy of data analysis in the area of SSP. In particular, the HSA algorithm [131] was implemented on filler sounds of female speakers and features extracted from the resulting Hilbert spectrum (called HS features hereafter) were then fed to a Support Vector Machine (SVM) classifier to perform personality analysis. The performance of the SVM classifier using the HS features was compared with that of using features extracted from short time Fourier transform (STFT), called STFT features hereafter. The results suggested that HS features are competitive with STFT features in terms of personality prediction accuracy and, thus, show the effectiveness of HSA for signal processing. However, the experiments showed that HSA produces a different number of features for every sample and hence it cannot be used easily in machine learning approaches.

The rest of this chapter is organised as follows: Section 7.1 describes the background of HSA, Section 7.2 presents the HSA algorithm, Section 7.3 reports on the experiments and results and Section 7.4 offers some conclusions.

7.1 Background

HSA generalises the definition of the Hilbert spectrum by using a superposition of complex AM-FM components parametrised by the IA and IF [131]. It is a type of time-frequency analysis of non-stationary signals, compared to the Fourier spectral analysis that assumes the system is linear and the data is strictly stationary [69]. While the Fourier transform outputs the energy-frequency distribution of the data, the output of HSA gives a full energy-frequency-time distribution of the data characterised by the IA and IF which will be discussed next.

Many physical phenomena are characterised by a complex signal

$$z(t) = x(t) + jy(t) = \rho(t)e^{j\Theta(t)} \tag{7.1}$$

where $\rho(t)$ is called the signal's IA, $\Theta(t)$ the signal's instantaneous phase (IP), and $\Omega(t) = \frac{d}{dt}\Theta(t)$ the signal's IF. $z(t)$ is in general called the latent signal because only the real part $x(t)$ is observed and the imaginary part $y(t)$ is hidden; that is, the observation is related to the real operator:

$$x(t) = \Re\{z(t)\} \tag{7.2}$$

Therefore, the complex extension problem, that is, determining $z(t)$ from $x(t)$ is called the latent signal analysis (LSA) problem. In the classical approach, a unique rule $\mathcal{L}\{.\}$ is sought to get an estimate of $y(t)$:

$$\hat{y}(t) = \mathcal{L}\{x(t)\} \tag{7.3}$$

This results in the instantaneous estimates:

$$\hat{\rho}(t) = \pm|z(t)| = \pm\sqrt{x^2(t) + \hat{y}^2(t)} \tag{7.4}$$

$$\hat{\Omega}(t) = \frac{d}{dt}\Big[\arctan\Big(\frac{\hat{y}(t)}{x(t)}\Big)\Big] \tag{7.5}$$

The Hilbert transform (HT) is almost universally the chosen rule; however, Sandoval et al. [131] argued that it limits one to only a subset of latent signals that have the same real part $x(t)$. After a review of the HT and motivations behind its use and the analytical signal, they proved that analyticity can still be maintained without HC and that no universal rule for the quadrature exists. The following Section 7.1.1 briefly reviews the HT and the analytical signal, and the reasons for relaxing the HC condition are given in Section 7.1.2.

7.1.1 The HT and Analytical Signal

The HT is defined as:

$$\mathcal{H}\{x(t)\} \equiv -\frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{x(\tau)}{\tau - t} d\tau \tag{7.6}$$

The adoption of the HT has three main motivations: 1) Vakman's physical conditions, 2) analyticity of the resulting complex signal, and 3) computational ease through Gabor's quadrature method (QM) [48]. The following reviews these motivations.

Vakman's Physical Conditions

Vakman [146, 147, 148, 149, 150, 151] proposed conditions to restrain the ambiguities in choosing the complex extension.

The first ambiguity arises when conducting instantaneous parametrisation for real signals [11, 113]. Consider

$$x(t) = \Re\Bigg\{a_0(t)e^{j[\int_{-\infty}^{t}\omega_0(\tau)d\tau + \phi_0]}\Bigg\} \tag{7.7}$$

There is an infinite set of pairs $a_0(t), \omega_0(t)$ corresponding to the same real signal $x(t)$, and

hence an infinite set of pairs IA, IF. To ensure that a unique complex extension can be justified for a real signal, Vakman proposed three conditions:

Condition 1: Amplitude Continuity: A small change in the real signal $x(t)$ should cause a correspondingly small change in IA $\rho(t)$. That is , $\rho(t)$ is a continuous function, which implies the rule in eq. (7.3) is continuous, i.e.

$$\mathcal{L}\{x(t) + \epsilon w(t)\} \to \mathcal{L}\{x(t)\} \quad \text{for} \quad ||\epsilon w(t)|| \to 0 \tag{7.8}$$

Condition 2: Phase Independence of Scaling and Homogeneity: Scaling the real signal $x(t)$ by a real constant c (> 0) should have no affect on IP $\Theta(t)$ and IF $\Omega(t)$, and IA $\rho(t)$ should be multiplied by c. This implies the complex extension rule in eq. (7.3) is scalable:

$$\mathcal{L}\{cx(t)\} = c\mathcal{L}\{x(t)\} \tag{7.9}$$

Condition 3: Harmonic Correspondence: Let $x(t) = a_0 \cos(\omega_0 t + \phi_0)$, that is, a pure sinusoid with constant amplitude a_0 and frequency ω_0. Then HC will lead to the complex extension:

$$\hat{z}(t) = a_0 e^{j(\omega_0 t + \phi_0)} \tag{7.10}$$

where the IA, IF are equal to the constants a_0 and ω_0, respectively. This implies that for any constant $a_0 > 0, \omega_0 > 0$, we must have:

$$\mathcal{L}\{a_0 \cos(\omega_0 t + \phi_0)\} = a_0 \sin(\omega_0 t + \phi_0) \tag{7.11}$$

and hence $\hat{z}(t)$ is a simple harmonic component (SHC) with positive IA and IF.

Vakman showed in [146, 149] that the HT is the only operator that meets the above conditions. Consequently, the HT (or Gabor's practical QM) is regarded as the correct way to complex extend a signal.

The following Condition (4) aims to resolve the ambiguity in complex signals: we construct $\hat{z(t)}$ using eq. (7.3), but we want the pair IA, IF for signal analysis, and ambiguities can occur in this coordinate transformation [49]. For example, there may be negative IA in eq. (7.4).

Condition 4: Phase continuity: The phase $\Omega(t)$ is a continuous function.

Gabor's QM

The HT is an ideal operator and hence is not physically realisable in practice. Gabor's QM [48] is a frequency domain method that is equivalent to the HT under certain conditions. It involves the following three steps:

1. decompose $x(t)$ into SHCs by computing its Fourier spectrum
2. double the magnitude of the non-negative frequency components
3. negate the negative frequency component

The resulting complex signal is formulated in terms of SHCs each having constant IA and non-negative IF. Gabor's QM is widely used because of its convenient practical implementation using Fourier transform (FT).

Analyticity of the Analytic Signal

The analytic signal (AS) refers to the complex signal resulting from the HT (i.e. Gabor's QM) of a real signal [6, 12, 20, 156, 167]. Define:

$$z(\iota) = u(t,\tau) + jv(t,\tau) \quad \text{with} \quad \iota \equiv t + j\tau \tag{7.12}$$

Then $z(\iota)$ is called an *analytic function* if its real and imaginary parts satisfy the Cauchy-Riemann (CR) conditions:

$$\frac{\partial}{\partial t}u(t,\tau) = \frac{\partial}{\partial \tau}v(t,\tau) \quad \text{and} \quad \frac{\partial}{\partial \tau}u(t,\tau) = -\frac{\partial}{\partial t}v(t,\tau) \tag{7.13}$$

For the AS obtained by the HT, that is, $\hat{z}(t) = x(t) + j\mathcal{H}\{x(t)\}$, if we set t to $\iota \equiv t + j\tau$, then the resulting complex function

$$\hat{z}(\iota) = \hat{u}(t,\tau) + j\hat{v}(t,\tau) \tag{7.14}$$

is an analytic function [13]. For this reason, the HT-extended complex signal $\hat{z}(t) = x(t) + j\mathcal{H}\{x(t)\}$ is called the AS [20, 156].

7.1.2 Relaxing the Condition of HC

The motivation for Vakman's HC assumption is that the SHC is the solution to the following differential equation:

$$\frac{d^2}{dt^2}z(t) + \omega_0^2 z(t) = 0 \tag{7.15}$$

which describes many ideal physical systems such a LC circuit or spring/mass model.

However, any modification to the above differential equation eq. (7.15) means deviation from the ideal case. Consider:

$$\frac{d^2}{dt^2}z(t) + c\frac{d}{dt}z(t) + \omega_0^2 z(t) = 0 \tag{7.16}$$

where c is a constant. The solution to eq. (7.16) is:

$$z(t) = a_0 e^{-vt} e^{j(\omega_d t + \phi_0)} \tag{7.17}$$

which includes an AM term and is not a SHC [76]. In addition, in cases where the differential equation contains time-varying coefficients or partial derivatives with respect to τ, the associated solution may consist of both AM and FM terms, which is not a SHC as well [37, 116, 150].

For these reasons, Sandoval et al. [131] advocate that the condition of HC is overly restrictive and can lead to incorrect interpretations because it is common that real physical systems deviate from the ideal one such as described in eq. (7.15). Consequently, they claim that, "*by not assuming HC, we gain a degree of freedom in our analysis that allows us to construct other complex extensions that may be better suited to describing the underlying physical phenomena associated with the signal*" and believe that, any attempt to find a unique rule to infer $z(t)$ defined in eq. (7.1) from $x(t)$ defined in eq. (7.2), is "*fundamentally flawed and that no such universal rule can exist*".

Following this claim, they proved the proposed Theorem "*If we do not assume HC, there exists at least one choice for the quadrature, $y(t) \neq \mathcal{H}\{x(t)\}$ that results in $z(t)$ being an*

analytic function " and the Corollary "*No unique rule for the quadrature,* $\hat{y}(t) = \mathcal{L}\{x(t)\}$ *exists to obtain the latent signal,* $z(t)$ *from the observation,* $x(t) = \Re\{z(t)\}$ *for all* $z(t)$."

Consequently, they argued that the HT (which assumes HC) can be used to determine the IA/IF only for a small subset of signals (those with HC), and alternative latent signals can be obtained by relaxing the HC condition. Moreover, they pointed out that Gabor's practical QM implementation (equivalent to the HT) can only be used in situations where a latent signal is a superposition of SHCs with non-negative IF and in cases where $\mathcal{H}\{x(t)\} \approx y(t)$ [113], e.g. communication signals that can approximately satisfy Bedrosian's theorem [7, 105, 150]. For non-linear and non-stationary signals, Gabor's complex extension fails to provide the necessary flexibility. To solve this problem, they proposed the Hilbert spectral analysis (HSA) employing a superposition of latent AM-FM components (rather than SHCs) which will be presented in the following section.

7.1.3 HSA Using Latent AM-FM Components

By relaxing the HC condition, Sandoval et al. [131] proposed the construction of latent signal $z(t)$ as a multi-component model comprising a superposition of K (possibly infinite) complex AM-FM components:

$$z(t) \equiv \sum_{k=0}^{K-1} \psi(t; a_k(t), \omega_k(t), \phi_k) \tag{7.18}$$

where the AM-FM component is defined by:

$$\begin{aligned} \psi(t; a_k(t), \omega_k(t), \phi_k) &\equiv a_k(t) \exp\left\{ j \left[\int_{-\infty}^{t} \omega_k(\tau) d\tau + \phi_k \right] \right\} \\ &= a_k(t) e^{j\theta_k(t)} \\ &= s_k(t) + j\sigma_k(t) \end{aligned} \tag{7.19}$$

with $a_k(t), \omega_k(t)$ representing the IA and IF for component k respectively, and ϕ_k being the phase reference. The observed signal $x(t)$ is associated with $z(t)$ according to eq. (7.2).

Examples of the above complex AM-FM components (eq. (7.19)) include constant ampli-

tude and constant frequency, time-varying frequency and constant amplitude, time-varying amplitude and constant frequency. As can be seen, SHC (constant amplitude and constant frequency) is just one special case of the AM-FM components, and relaxation of HC allows greater freedom in the construction of the latent signal. In this case, the complex extension is implied by the assumptions of the underlying signal model. Consequently, correct assumptions must be made on a per problem basis in order to arrive at a unique decomposition and hence a proper estimation of the latent signal and its components. The resulting Hilbert spectrum is parametrised by a set of IA/IF pairs, each associated with the corresponding component.

Although this newly proposed multi-component representation of the signal (termed as HSA by [131]) tends to be a powerful signal analysis technique theoretically, its application in signal processing has not been explored. This chapter aims to bridge this gap and attempts to explore the usefulness of HSA in practical applications. In particular, the extraction of features from the Hilbert spectrum (HS) of the filler sounds of female speakers will be explored, and the resulting HS features will be used to perform a similar personality analysis to that described in detail in Chapter 6. In the following section, the HSA algorithm used in this work will be presented.

7.2 HSA Algorithm

The HSA algorithm adopted in this work is called the HSA-IMF algorithm, the pseudo codes of which can be found in Algorithm 9 in [131]. This algorithm builds upon the empirical mode decomposition (EMD) of [69], ensemble EMD (EEMD) of [166], complete EEMD of [144] and tone masking of [29], but proposes the intrinsic mode function (IMF) demodulation rather than the original HT demodulation in EMD of [69] to obtain the estimates of IA and IF. The reason is that, under the HSA framework, HT is no longer appropriate for the demodulation since HT assumes SHCs and HC whereas IMF is generally a latent AM-FM component without HC. The EMD of [69] sequentially found a set of AM-FM components (IMFs) through a sifting algorithm which iteratively identifies and removes the trend from the signal (i.e. behaves as a high pass filter). EEMD and tone masking admitted the ensemble averaging to solve the mode mixing problem caused by the signal intermittency (i.e.

the relative component intermittency). The complete EEMD aimed to address some of the undesirable characteristics of EEMD such as high computational complexity, the loss of the perfect reconstruction property and propagation of IMF estimation error. The way the complete EEMD works is that it averages at the component-level when estimating each component rather than the average at the conclusion of all EMD trials, and hence requires fewer sifting iterations, a smaller ensemble size and is able to recover the completeness property of the original EMD algorithm. The tone masking method differs from the EEMD by using a deterministic (masking) signal rather than a noise signal to help track the components properly, and can perform better with a carefully chosen masking signal. This HSA-IMF algorithm incorporates the most desirable features of the complete EEMD and tone masking to solve the mode mixing problem, Rato's improvements to the sifting algorithm [122] and Sandoval et al.'s self-proposed IMF demodulation method [131]. The following section presents the specific parameter settings for the HSA-IMF algorithm used in the experiments in this work (see Section 7.3).

7.2.1 Parameter Settings for the HSA-IMF

Table 7.1 shows the specific experimental settings for the HSA-IMF algorithm used in this work - the "HSAr2" toolkit developed by [131].

Table 7.1 Experimental settings for the HSA-IMF algorithm.

SiftStopThresh	Sifting stop criterion in dB (usually start around 30dB)	30
EMDStopThresh	EMD stop criterion in dB (usually start around 10dB)	10
alpha	sifting step size (normally<=1)	0.95
I	number of ensemble trials	2
beta	scale factor for noise	$[0.01, 2]$

The "HSAr2" toolkit was implemented to output the Hilbert spectrum (HS) for the data set presented in the following Section 7.3 . The HS output includes a set of instantaneous amplitude (IA) and instantaneous frequency (IF) for the latent AM-FM components decomposed. Experimental exploration of the application of HSA-IMF algorithm - the experiments of extracting features from the resulting IAs and IFs - will be reported in the following sections.

7.3 Experiments and Results

In this section, we investigate feature extraction from the Hilbert spectrum (HS) of the voice data. The data set used are the fillers of female speakers, and the total number of data samples was 716. Similar to Chapter 6, the Big-Five personality traits [132] were attributed by experts to each female speaker according to her filler sound. In order to test the effectiveness of the resulting HS features, the classification accuracy of the SVM classifier (see Section 6.5.3) using HS features is compared with that of using the feature set extracted from the same data based on short time Fourier transform (STFT). The reason for choosing a SVM classifier rather than a Gaussian process classifier (described in Chapter 4) is that here we are just exploring features and we are not too concerned with the quantification of uncertainty. The STFT features are obtained using openSmile - the state-of-the-art audio feature extraction toolkit based on the STFT output. The classification task is to predict whether a female speaker is perceived to be above the median or not with respect to each of the Big-Five personality traits from her filler sounds on a voice call. The Big-Five traits are: openness, conscientiousness, extraversion, agreeableness and neuroticism as described in Section 6.3.2, denoted by "Ope","Con","Ext", "Agr" and "Neu" respectively in the following sections. The prediction accuracies result from a ten-fold cross-validation approach. In the following section, feature extraction using openSmile on the fillers data will be introduced.

7.3.1 Feature Set From openSmile

The INTERSPEECH 2009 Emotion Challenge feature set (384 features) was selected using openSmile [36] as it was more suited to the personality analysis in this case. The 384 features are generated by applying twelve statistical functionals (shown in Table 7.2) to sixteen smoothed low-level descriptor contours as well as their 1st order delta coefficient (differential). The sixteen low-level descriptors were as follows:

1. Zero-crossing-rate (ZCR) of the time signal
2. Root mean square (RMS) signal frame energy
3. The fundamental frequency computed from the Cepstrum

4. The voicing probability (harmonics-to-noise ratio) computed from the ACF (autocorrelation function)

5. Mel-frequency cepstral coefficients (MFCC) one to twelve

Table 7.2 12 functionals for extracting features for emotion recognition as analysed by [36].

max	The maximum value of the contour
min	The minimum value of the contour
range	max - min
maxPos	The absolute position of the maximum value (in frames)
minPos	The absolute position of the minimum value (in frames)
amean	The arithmetic mean of the contour
linregc1	The slope (m) of a **linear approximation** of the contour
linregc2	The offset (t) of a **linear approximation** of the contour
linregerrQ	The quadratic error computed as the difference of the **linear approximation** and the actual contour
stddev	The standard deviation of the values in the contour
skewness	The skewness (3rd order moment)
kurtosis	The kurtosis (4th order moment)

In the next section, the extraction of features from the HS of the fillers data will be presented.

7.3.2 Feature Extraction From the HS

This section presents the experiments on feature extraction from the IA and IF of the Hilbert spectrum (HS) of the fillers data. The HS of the data was obtained by implementing the HSA-IMF algorithm (experimental settings of which can be found in Section 7.2) on the complete 716 data samples. Since each data sample may have a different number of underlying components because of the nature of the HSA-IMF approach, it is difficult to derive the same number of features for all the 716 data instances. Consequently, the following settings for extracting features from the HS of the data were explored. It should be noted that, whenever a t-test is mentioned in the following settings, it means one-sided t-test for the results obtained from ten repeated experiments.

Setting I

In this first setting, only data that have six or seven intrinsic mode functions (IMFs) were selected, resulting in a total number of 686 samples (i.e. 96% of the data have six or seven IMFs). Next, six statistical functionals (mean, median, standard deviation, maximum, minum, range) were applied to each pair of the IA and IF of the last five IMFs. Apart from the purpose of obtaining the same number of features for each data instance, another reason for choosing the last five IMFs is that the first IMF in most cases is noise and this does not give much information. This leads to sixty ($6 \times 5 \times 2$) features for each data instance. This case is denoted by HS_60. The binary classification accuracy of SVM using HS_60 features is compared with that of exploiting 384 features extracted using openSmile (denoted by OS_384) in Table 7.3. As can be seen, the classification accuracies with only 60 features from the HS are well above chance for all traits.

Table 7.3 Prediction accuracy of HS_60 and OS_384_I for each of the Big-Five traits.

TRAIT	HS_60	OS_384_I
Ope	0.69	0.82
Con	0.62	0.77
Ext	0.62	0.74
Agr	0.72	0.82
Neu	0.63	0.73

HS_60 denotes the case where the prediction accuracy of SVM is obtained using the 60 features extracted by applying 6 statistical functionals to each pair of the IA and IF of the last 5 IMFs resulting from the Hilbert spectrum of the 686 data instances that have 6 or 7 IMFs.
OS_384_I denotes the case where the prediction accuracy of SVM is obtained using the 384 features extracted from the 686 data instances via openSmile.

Setting II

In this setting, 553 data instances that have seven IMFs were selected. The twelve statistical functionals listed in Table 7.2 were then applied to each pair of the IA and IF of the 7 IMFs, resulting in 168 ($12 \times 7 \times 2$) features for each data instances. This case is denoted by HS_168. The linear approximation highlighted in Table 7.2 is performed between the IA or IF and instantaneous time. In Table 7.4, the classification accuracy of SVM using HS_168

features is compared with that of exploiting 384 features extracted via openSmile (denoted by OS_384_II in this case). The absolute value of the difference between the prediction accuracy of the HS_168 case and the OS_384_II case is denoted by Abs_diff_II. Column 4 of Table 7.4 shows the results from the one-sided t-test for the values of Abs_diff_II for the five traits with p-values shown in brackets. The results suggested that raw features from the HS (only 168) are capable of achieving comparable prediction accuracy with 384 openSmile features that have been extracted with very sophisticated methods (p-value < 0.025). In addition, the much higher prediction accuracy (one-sided t-test, p-value < 0.025) of the HS_168 case than that of the HS_60 case (in Table 7.3) seems to suggest that the application of twelve functionals to get more features does perform much better than applying six functionals to the HS of the data.

Table 7.4 Prediction accuracy of HS_168 and OS_384_II for each of the Big-Five traits.

TRAIT	HS_168	OS_384_II	Abs_diff_II (p-value)
Ope	0.79	0.84	<0.05 (5.94e-07)
Con	0.68	0.79	<0.1 (1.12e-03)
Ext	0.69	0.72	<0.05 (1.09e-03)
Agr	0.75	0.82	<0.06 (5.38e-04)
Neu	0.73	0.75	<0.03 (3.34e-03)

HS_168 denotes the case where the prediction accuracy of SVM is obtained using the 168 features extracted by applying 12 statistical functionals to each pair of the IA and IF of the 7 IMFs of the 553 data instances that have 7 IMFs. OS_384_II denotes the case where the prediction accuracy of SVM is obtained using the 384 features extracted from the 553 data instances via openSmile. Abs_diff_II denotes the absolute value of the difference between the prediction accuracies of the HS_168 case and the OS_384_II case. The p-values are obtained from one-sided t-test.

Setting III

In this setting, for the 553 data instances that have seven IMFs, the 144 MFCC features of the 384 openSmile features are replaced with the 168 HS features described in Setting II. This leads to 408 (384 - 144 + 168) features for each data instance. This case is denoted by HS_OS_408. This is because, as the 144 MFCC features and the 168 HS features were obtained by applying the same 12 functionals (see Table 7.2) to the smoothed STFT output and the HS output respectively, the performance of this combination of features could indi-

cate the predictive effect of the HS features. Table 7.5 displays the classification accuracy of HS_OS_408 and OS_384. The absolute value of the difference between the prediction accuracy of the HS_OS_408 case and the OS_384_II case is denoted by Abs_diff_III. The results from the one-sided t-test for the values of Abs_diff_III for the five traits are shown in Column 4 of Table 7.5, with p-vales shown in brackets. The results suggested that the HS_OS_408 case is able to achieve comparative prediction accuracy with that of the OS_384 case (p-vale < 0.025), and hence showed the effectiveness of the HS features as an alternative to the STFT features.

Table 7.5 Prediction accuracy of HS_OS_408 and OS_384_II for each of the Big-Five traits.

TRAIT	HS_OS_408	OS_384_II	Abs_diff_III (p-value)
Ope	0.80	0.84	<0.04 (3.03e-05)
Con	0.74	0.79	<0.05 (2.02e-04)
Ext	0.74	0.72	<0.02 (1.56e-02)
Agr	0.81	0.82	<0.02 (6.33e-04)
Neu	0.69	0.75	<0.06 (8.67e-04)

HS_OS_408 denotes the case where the prediction accuracy of SVM is achieved using the 408 features obtained by replacing 144 MFCC features of the 384 openSmile features with the 168 HS features.
OS_384_II denotes the case where the prediction accuracy of SVM is obtained using the 384 features extracted from the 553 data instances via openSmile.
Abs_diff_III denotes the absolute value of the difference between the prediction accuracies of the HS_OS_408 case and the OS_384_II case. The p-values are obtained from one-sided t-test.
Note: the data used for the two cases are the 553 data instances that have 7 IMFs.

Setting IV

In this setting, for the 553 data instances with seven IMFs, the twelve statistical functionals (see Table 7.2) were first applied to the IA and IF of the *superposition* of the seven IMFs. This leads to 24 (12×2) features for each data instance. This case is denoted by HS_SU_24. As in Setting III, the 144 MFCC features were also replaced with the 24 HS_SU_24 features. This results in 264 (384 - 144 + 24) features for each data instance. This case is denoted by HS_SU_264. Table 7.6 presents the classification accuracy of HS_SU_24, HS_168, HS_SU_264, HS_OS_408. The difference of the prediction accuracy between the HS_168 case and the HS_SU_24 case is denoted by Diff_IV_A (the former minus the lat-

ter), and that between the HS_OS_408 case and HS_SU_264 case is denoted by Diff_IV_B (the former minus the latter). The results of the one-sided t-test for the values of Diff_IV_A and Diff_IV_B are shown in Columns 4 and 7 of Table 7.6, respectively. As can be seen, the classification accuracies of the HS_SU_24 case reduced significantly compared to those of the HS_168 case (p-value < 0.025). This indicates that extracting features from the superposition of the IMFs does not seem to give much information. The reduced prediction accuracies of the HS_SU_264 case compared to those of the HS_OS_408 case (p-value < 0.025) also suggest that it is appropriate to extract features from every pair of IA and IF of all the IMFs rather than from one pair of IA and IF obtained from the superposition of all the IMFs.

Table 7.6 Prediction accuracy of HS_SU_24, HS_168, HS_SU_264 and HS_OS_408 for each of the Big-Five traits.

TRAIT	HS_SU_24	HS_168	Diff_IV_A (p-value)	HS_SU_264	HS_OS_408	Diff_IV_B (p-value)
Ope	0.68	0.79	>0.11 (1.01e-04)	0.75	0.80	>0.04 (4.97e-05)
Con	0.45	0.68	>0.20 (3.86e-03)	0.70	0.74	>0.04 (4.13e-03)
Ext	0.39	0.69	>0.27 (1.06e-02)	0.68	0.74	>0.05 (1.57e-04)
Agr	0.63	0.75	>0.12 (1.88e-04)	0.73	0.81	>0.07 (1.26e-03)
Neu	0.38	0.73	>0.29 (5.14e-03)	0.69	0.70	>0 (1.48e-02)

HS_SU_24 denotes the case where the prediction accuracy of SVM is achieved using the 24 features extracted by applying 12 statistical functionals to one pair of the IA and IF of the superposition of the 7 IMFs resulting from the Hilbert spectrum.
HS_168 denotes the case where the prediction accuracy of SVM is obtained using the 168 features extracted by applying 12 statistical functionals to each pair of the IA and IF of the 7 IMFs of the 553 data instances that have 7 IMFs.
HS_SU_264 denotes the case where the prediction accuracy of SVM is achieved using the 264 features obtained by replacing 144 MFCC features of the 384 openSmile features with the 24 HS features extracted from the superposition of the 7 IMFs.
HS_OS_408 denotes the case where the prediction accuracy of SVM is achieved using the 408 features obtained by replacing 144 MFCC features of the 384 openSmile features with the 168 HS features extracted from the 7 IMFs directly.
Diff_IV_A denotes the difference of the prediction accuracies between the HS_168 case and the HS_SU_24 case (the former minus the latter); Diff_IV_B represents the difference of the prediction accuracies between the HS_OS_408 case and HS_SU_264 case (the former minus the latter). The p-values are obtained from the the one-sided t-test.
Note: the data used for the four cases are the 553 data instances that have 7 IMFs.

Setting V

As mentioned earlier, the nature of HSA means each data instance may have a different number of IMFs. Therefore, this makes it difficult to generate the same number of features for all data instances. To solve this problem , only data instances that have six or seven IMFs were selected in Settings I - IV. However, this means that it is not possible to use all the data when making predictions. This setting V attempts to make sure all the 716 data instances have the same number of features in the following way. First, the Hilbert spectrum of data instances that have fewer than seven IMFs is padded with zeros (e.g. if the number of IMFs is five, then the first two IMFs are padded with zeros). Next, in data instances for which the number of IMFs is greater than seven only the last seven IMFs are selected. This will result in the same number of IMFs and hence the same number of features (168) for all data. The reason for choosing seven IMFs is that 60% of all data have seven IMFs. This case is denoted by HS_PAD_168. Table 7.7 shows the classification accuracy of HS_PAD_168 and HS_168. The difference between the prediction accuracies of the HS_168 case and the HS_PAD_168 case is denoted by Diff_V (the former minus the latter). Column 4 of Table 7.7 displays the results from the one-sided t-test for the values of Diff_V for the five traits, with p-values shown in brackets. It can be seen that the prediction accuracies of the HS_PAD_168 case drop a lot compared to those of the HS_168 case (p-value <0.025). Therefore, using padding to produce the same number of IMFs and hence to obtain the same number of features for all data, does not seem to be feasible for the application of Hilbert spectral analysis.

7.4 Conclusions

This chapter has presented the Hilbert spectral analysis (HSA) framework proposed by [131] and the experimental exploration of the extraction of features from the Hilbert spectrum (HS) of fillers of female speakers. This has been the first attempt to investigate the application of HSA-IMF in signal processing. Five settings for extracting features from the HS have been explored. The resulting HS features were fed into a SVM classifier to discriminate between female speakers scoring above and below the median of the scores for all Big-Five traits. The prediction accuracy of SVM using HS features has been compared with that of using

Table 7.7 Prediction accuracy of HS_PAD_168, HS_168 for each of the Big-Five traits.

TRAIT	HS_PAD_168	HS_168	Diff_V (p-value)
Ope	0.66	0.79	>0.12 (1.33e-03)
Con	0.55	0.68	>0.13 (1.82e-04)
Ext	0.62	0.69	>0.05 (4.49e-04)
Agr	0.69	0.75	>0.05 (6.41e-04)
Neu	0.56	0.73	>0.13 (1.75e-03)

HS_PAD_168 denotes the case where the prediction accuracy of SVM is obtained using the 168 features extracted by applying 12 statistical functionals to each pair of the IA and IF of the 7 IMFs of all the 716 data instances where the Hilbert spectrum of the data instances that have fewer than 7 IMFs is padded with 0s such that they also have 7 IMFs.
HS_168 denotes the case where the prediction accuracy of SVM is obtained using the 168 features extracted by applying 12 statistical functionals to each pair of the IA and IF of the 7 IMFs of the 553 data instances that have 7 IMFs.
Diff_V denotes the difference between the prediction accuracies of the HS_168 case and the HS_PAD_168 case (the former minus the latter). The p-values are obtained from the one-sided t-test.

features extracted from the STFT output. The results suggested that exploiting HS features achieves competitive prediction accuracy with using STFT features, and hence showed the effectiveness of HSA in practical applications and indicated an alternative feature extraction method for signal processing. However, the variability of the number of IMFs in HSA makes it difficult to generate the same number of features for all data. As a fixed feature size is expected in most cases in machine learning, finding a good solution for seamlessly using machine learning on the computed Hilbert spectrum could be one direction for future work.

Chapter 8

Conclusions

Gaussian processes have been widely applied in a number of domains. The popularity of GP models is largely attributed to their attractive properties: the probabilistic non-parametric formulation makes them capable of quantifying uncertainty while offering flexibility when modelling data and the parametrisation in the covariance function makes it possible to gain insights into the application under study. These properties are particularly desirable in the area of social signal processing, where the inference techniques are dominated by non-probabilistic parametric models.

However, the ability of the GP models to quantify uncertainty relies on an accurate characterisation of the posterior distribution with respect to the covariance parameters. This is normally done by means of standard Markov chain Monte Carlo (MCMC) algorithms, which suffer from the problem of inefficiencies: their rejection of proposals wastes a considerable amount of expensive calculations and in the case of pseudo-marginal MCMC, an overestimation of the marginal likelihood may cause the chain getting stuck in certain regions of the space.

In this thesis, an alternative inference framework for GP models based on the adaptive multiple importance sampling (AMIS) has been proposed in order to overcome the foregoing limitations of MCMC approaches. There has been an extensive comparison of the sampling efficiency of MCMC versus AMIS for GP models, and the results showed that AMIS can achieve faster convergence speed for both scenarios of GP regression and classification. The

GP-based pseudo-marginal AMIS with a newly designed kernel (G-ARD) to perform personality inference - an important area of social signal processing - was also applied. The task was to classify personality traits of people from the online Flickr pictures. The results demonstrated the value of the proposed GP framework for social signal processing: apart from achieving high prediction accuracies, the G-ARD kernel is capable of identifying features that better account for the classification outcome, information from which can be used to change our impressions online.

There was also an attempt, possibly for the first time in the literature, to investigate feature extraction under the Hilbert spectral analysis framework. The results indicated that, for social signal processing, features extracted from the Hilbert spectrum can work as an effective alternative to features extracted from the short time Fourier transform output.

The rest of this chapter is organised as follows. Section 8.1 summarises the work reported in this thesis , Section 8.2 discusses limitations of the research as well as prospective future research and Section 8.3 ends with the contributions of this thesis.

8.1 Thesis Summary

As Chapter 1 presented the introduction of the thesis, and Chapters 2 and 3 reviewed the Bayesian Gaussian processes and MCMC approaches respectively, the thesis summary in this section focuses on Chapters 4 to 7, the contents of which have been the main contributions of this thesis.

Chapter 4, *Adaptive Monte Carlo*, examined AMIS for GPs with Gaussian likelihoods and proposed PM-AMIS for those with non-Gaussian likelihoods. The motivation was as follows. Inference of GP covariance is usually costly because it requires repeatedly calculating the marginal likelihood. Even in the case of Gaussian likelihood where the marginal likelihood is computable, the computation of the marginal likelihood is expensive as it has time complexity scaling with the cube of the number of input data instances. In the non-Gaussian likelihood case, using approximations such as LA or EP to approximate the marginal likelihood further increases the time complexity. Most of the literature proposes the use of MCMC to characterise the posterior distribution over GP covariance parameters in order to quantify

uncertainty without introducing any bias and, although MCMC proved successful in various scenarios, employing MCMC algorithms may result in inefficiencies for the following reasons:

- MCMC methods are based on the iteration of the following two operations: (i) proposal and (ii) an accept/reject step. For instance, optimal acceptance rates for popular MH and HMC are approximately 25% and 65%, respectively. Given that calculating the marginal likelihood and possibly the gradients of its logarithm with respect to the covariance parameters is expensive, whenever MCMC algorithms reject a proposal, a considerable amount of computations are therefore wasted.

- In GPs with non-Gaussian likelihood, the PM-MCMC approach bypasses the need to compute the marginal likelihood exactly, but may suffer from inefficiencies because when a proposal is accepted and the marginal likelihood is largely overestimated, it becomes difficult for the chain to accept any other proposal.

Consequently, this thesis has proposed AMIS to alleviate the sampling inefficiencies of MCMC when inferring the GP covariance parameters.

The motivation of PM-AMIS was that it is impossible to compute exactly the marginal likelihood in the case of non-Gaussian likelihood, and hence an unbiased estimation of the likelihood was needed. A theoretical analysis was provided showing under which conditions the proposed PM-AMIS produces expectations under the posterior over GP covariance parameters without introducing any bias.

The contribution of this chapter was in proposing AMIS for GPs and in proposing PM-AMIS together with a theoretical analysis of it, which makes it possible to apply AMIS to any likelihood (Gaussian or non-Gaussian).

Chapter 5, *Experiments and Results*, provided an extensive comparison of convergence speed with respect to the computational complexity of AMIS versus MCMC.

The experiments were conducted for both the GP regression and classification cases, both of which employed six benchmark data sets. The number of data ranged from 214 to 1030, and the maximum number of dimensions was 20. The results suggested that importance

sampling-based inference of GP covariance parameters is competitive with MCMC algorithms; it is possible to achieve convergence of expectations under the posterior distribution of covariance parameters faster than employing MCMC methods in a wide range of scenarios. Even in the case of around twenty parameters, where importance sampling-based methods start to degrade in performance, this proposal is still competitive with MCMC approaches.

The contribution of this chapter was in the extensive comparison of convergence speed with respect to the computational complexity of AMIS versus MCMC and in empirically showing that AMIS is an effective alternative to MCMC when inferring GP covariance parameters. The extensiveness of the experiments was evidenced by the following:

1. six benchmark data sets were tested for both the GP regression and classification cases, each making use of both the RBF and ARD kernels;
2. for AMIS/MAMIS, importance distributions with two types of covariance matrices as well as regularised importance distribution were examined;
3. for MCMC algorithms, state-of-the-art MH/PM-MH, HMC (including NUTS and NUTSDA) and SS were considered;
4. for MH and HMC/NUTS/NUTSDA, three kinds of covariances of the starting proposals were investigated for tuning;
5. for PM-AMIS/PM-MH, both the LA and EP approximations were exploited.

Chapter 6, *Gaussian Processes for Finding Difference Makers in Personality Impressions - an Application of PM-AMIS*, investigated the application of PM-AMIS to SSP. Using features of visual characteristics extracted from the Flickr images (Section 6.4), the task was to predict whether a Flickr user is perceived to be above the median of the scores with respect to each of the Big-Five traits. The motivation was as follows. Most of the literature for personality inference (Section 6.2) has focused on non-probabilistic parametric methods that fall short of taking into account the uncertainty in the data or on non-parametric models such as SVM that gives no indication about the predictive effect of each feature. The experiments in this chapter attempted to exploit non-parametric Gaussian processes models (PM-AMIS

with a newly designed G-ARD kernel) to map favourite pictures into personality traits. The prediction accuracies achieved with PM-AMIS were compared with those achieved with SVM. The results suggested that, the new G-ARD kernel designed for the experiments of this chapter allowed a GP-based classifier (PM-AMIS) to achieve comparable accuracies as state-of-the-art SVM. Furthermore, the parameters of the G-ARD kernel allowed the identification of the groups of features that better account for the classification outcome while detecting cultural differences between UK and Asian personality assessors.

The contribution of this chapter was, firstly, in that it is the first work that has employed GPs with PM-AMIS, a fully probabilistic and non-parametric approach, to infer personality traits from online pictures and, secondly, in proposing the G-ARD kernel that learns automatically the interplay between groups of features and personality factors during the training process and provides indications about cultural effects through its weight differences.

Chapter 7, *Feature Extraction Using Hilbert Spectral Analysis*, provided an experimental investigation of feature extraction from the Hilbert spectrum. The motivation was as follows. Feature extraction in signal processing is dominated by various methods based on short time Fourier transform (STFT). Hilbert spectral analysis (HSA), recently proposed by [131], offers a new representation of signal fundamentally different from STFT. Consequently, the goal of this chapter was to explore feature extraction from this newly proposed HSA and its application in social signal processing. Similar to chapter 6, the task is to predict whether a female speaker belongs to the class *high* (above the median of the scores) or *low* (below the median of the scores) for each of the Big-Five traits employing features extracted from the Hilbert spectrum of her filler sounds. The prediction accuracies achieved with the SVM classifier using HS features were compared with those achieved using STFT features. The results suggested that using HS features achieved competitive prediction accuracies with those achieved using STFT features, and hence showed the effectiveness of HSA for signal processing.

The contribution of this chapter was in demonstrating a practical application of HSA to social signal processing for the first time and in providing an alternative feature extraction method for social signal processing.

8.2 Future Work

There are many interesting directions for future work which could follow from the results presented in this thesis. The following sections will discuss these.

8.2.1 AMIS Under Sparse GPs

In Chapter 4, AMIS was proposed to avoid the sampling inefficiencies of the MCMC algorithms for the inference of GP covariance parameters and extensive experiments on the convergence analysis of AMIS versus MCMC were reported in Chapter 5. As analysed in Chapter 5 (Section 5.4), the computational complexity of GP models largely comes from the computation and inversion of the covariance matrix and in the case of non-Gaussian likelihood, the LA and EP approximation significantly further increases the computational complexity. These make the GP models scale poorly in the number of data and this limitation applies when using AMIS for GP models. A variety of sparse GP models have been proposed in the literature for efficient computation when the number of data is large. An early attempt of [25] proposed the strategy of retaining a subset of the data. An inducing point approach was introduced by [135], augmenting the model with additional variables. Titsias [143] exploited the ideas of [135] in a variational approach. Authors of [82, 136] have introduced approximations based on the spectrum of the GP. Inducing point schemes that assume a Gaussian posterior have been proposed by [16, 87, 64] and can reduce the required computation enormously. Inspired by the findings of [93] that the variational inducing point method can minimize the Kullback-Leibler divergence to the posterior process, recent work by [63] presented a general inference scheme that extends the variational inducing point framework and allows for non-Gaussian posteriors. It would be interesting to explore how AMIS works under the foregoing sparse GP frameworks in the future.

8.2.2 Parallelisation of AMIS for GPs

Apart from resorting to spare GPs to improve computational efficiency, it is possible to exploit the fact that importance sampling-based algorithms are inherently parallel. Conse-

quently, with the advances in techniques of GPU and distributed computing, parallelisation of AMIS for GPs to accelerate computations is possible, and hence is another useful area to investigate in the future.

8.2.3 Use of the G-ARD Kernel

It is acknowledged that importance sampling suffers from the curse of dimensionality - for large dimensional problem, its performance degrades dramatically, and this applies to AMIS as well. One way to mitigate this is to use the G-ARD kernel proposed in Chapter 6. By using the G-ARD covariance function, the number of dimensions can be reduced to a level at which AMIS still performs well while providing meaningful information about each group of features. Therefore, an interesting possibility for future work is to exploit the G-ARD kernel in AMIS for large dimensional problem.

In Chapter 6, GPs with the proposed PM-AMIS, were used to predict personality traits from online pictures. The novel G-ARD kernel was able to detect the groups of features that better account for the classification outcome. According to recent sociological investigations [119], this is important because the impression people convey online can change the outcome of important issues such as, e.g. getting or not getting a job [23]. For this reason, future work could concentrate on how to use the information provided by the G-ARD weights to ensure that items posted online do not convey a wrong impression, whether it comes to faves or other types of online material.

8.2.4 Seamlessly Using Machine Learning on the Computed Hilbert Spectrum

In Chapter 7, feature extraction from the Hilbert spectrum of fillers of female speakers was explored. The results of the experiments showed the usefulness of the Hilbert spectrum in practical applications and thus indicated an alternative feature extraction method for social signal processing. However, because of the nature of the HSA-IMF approach, every data sample may have a different number of underlying components leading to an inability to generate the same number of features for all data. Since a fixed feature size is expected in

most cases in machine learning, one direction of future work could focus on finding a good solution for seamlessly using machine learning on the computed Hilbert spectrum.

8.3 Final Remarks

It is a challenging task to infer GP covariance parameters and to effectively apply the GP models to the analysis of scenarios in social signal processing. This thesis has addressed the sampling efficiency problem of MCMC algorithms that are normally employed, when inferring GP covariance parameters, by providing an alternative sampling framework based on AMIS. This thesis has also showed that, for social signal processing, GPs with AMIS is a powerful Bayesian approach capable of quantifying uncertainty while providing meaningful insights about the features. For feature extraction in social signal processing, this thesis has demonstrated novel results showing that Hilbert spectral analysis can work as an effective alternative to short time Fourier transform.

With regards to the future, there are opportunities to investigate AMIS for spare GPs to deal with larger data sets and to parallelise AMIS to accelerate computations; it would be interesting to study the application of the G-ARD kernel to AMIS for large dimensional problems and to explore how to use the information provided by the G-ARD kernel to convey a correct impression through online materials; and there are also opportunities to investigate seamlessly using machine learning on the computed Hilbert spectrum.

However, despite these possible future investigations, this thesis has presented a compelling inference framework for GP models and has demonstrated its power of quantifying uncertainty and interpretability for analyses in social signal processing. During this thesis, the following novelties with respect to the state-of-the-art were developed: (i) the application of AMIS to infer GP covariance parameters with any likelihood; (ii) a theoretical analysis of PM-AMIS; (iii) an extensive comparison of convergence speed with respect to computational complexity of AMIS versus MCMC methods; (iv) an application of PM-AMIS with the novel G-ARD kernel to perform personality inference from online pictures that has demonstrated the value of the proposed non-parametric probabilistic framework for SSP; (v) the first experimental exploration of feature extraction from the Hilbert spectrum that provides

an alternative feature extraction method for social signal processing.

Appendix A

Convergence of samplers for GP regression

Appendix A shows the convergence results of AMIS family (AMIS/MAMIS and their variants), MH family (MH-I/MH-D/MH-H) and HMC family (standard HMC , NUTS, NUTSDA) for the three regression data sets (Concrete, Housing and Parkinsons). Figures A.1 to A.3 show the convergence results for the samplers with the RBF covariance (RBF covariance case), whereas Figures A.4 to A.6 are convergence results for those with the ARD covariance (ARD covariance case).

Sub-figure (a) of Figures A.1 to A.6 demonstrate the results of AMIS/MAMIS. It can be seen that AMIS/MAMIS that exploits the full covariance structure of the proposal distribution performs better than the one that only updates the diagonal of the covariance matrix of the proposal density.

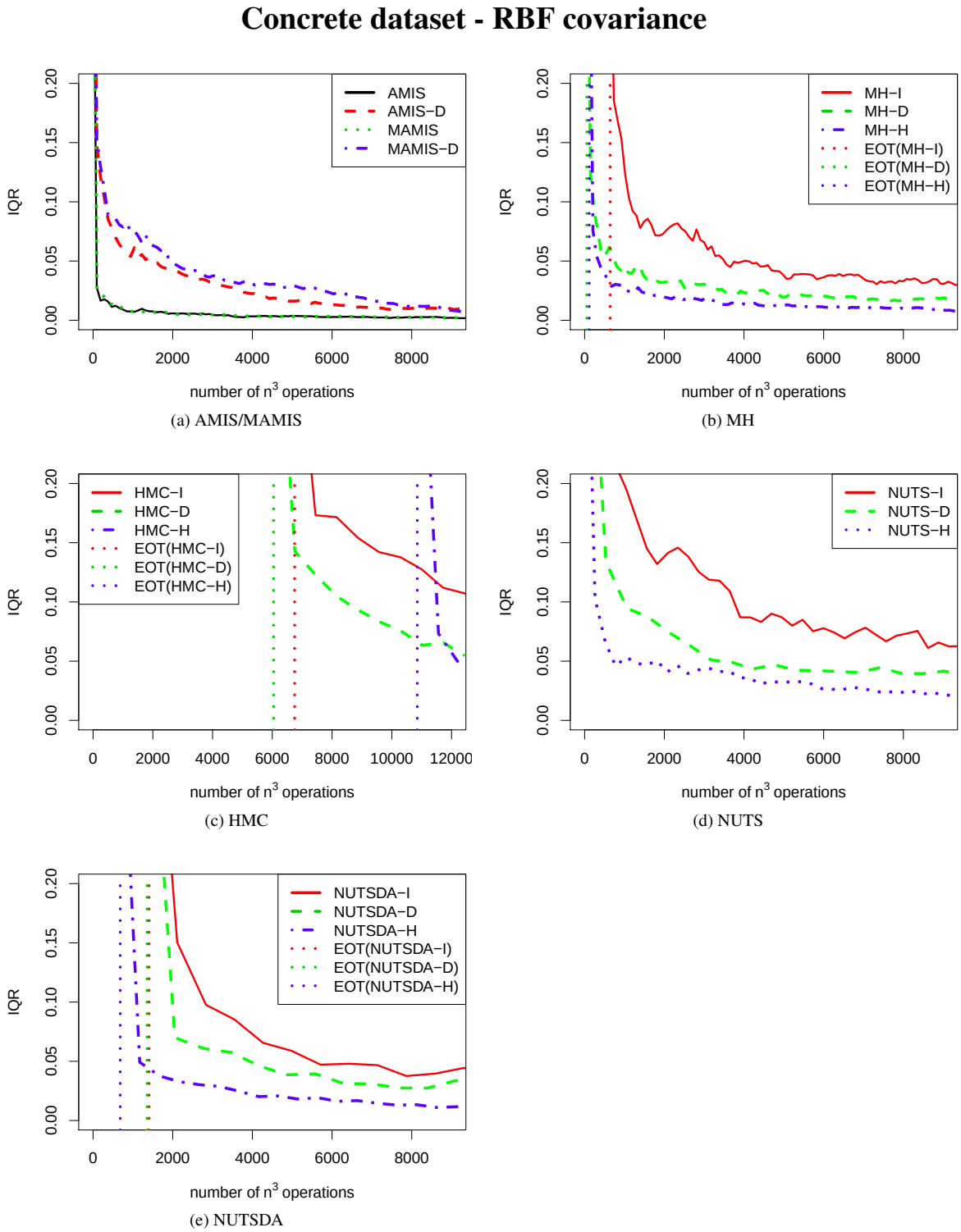
For the MH family and HMC family, sub-figures (b), (c), (d), (e) of Figures A.1 to A.6 show that, the methods that make use of the scales and correlation of the parameters, perform better than the ones that do not in most cases. Also, NUTS/NUTSDA turns out to converge much faster than the standard HMC because standard HMC has to be tuned costly in pilot runs. For MH and standard HMC, the computational cost of tuning is counted when comparing the convergence, as is shown in sub-figures (b), (c) of Figures A.1 to A.6 where the end of tuning (EOT) is indicated by three vertical dotted lines, corresponding to the three

variants respectively from left to right. For NUTSDA, the computational cost of tuning the parameters of the dual averaging scheme is also counted when determining the convergence, as is displayed in sub-figure (e) of Figures A.1 to A.6 with EOT indicated by three vertical dotted lines, relating to the three variants respectively from left to right. Table A.1 shows the corresponding computational cost of tuning:

Table A.1 Computational cost of tuning for HMC/NUTSDA.

	Concrete		Housing		Parkinsons	
	RBF	ARD	RBF	ARD	RBF	ARD
HMC-I	6.75	5.91	4.78	3.92	1.56	1.34
HMC-D	6.04	7.32	7.28	7.73	8.88	8.47
HMC-H	10.85	9.45	10.99	8.86	10.87	8.74
NUTDA-I	1.40	3.53	1.19	7.43	1.34	6.49
NUTDA-D	1.36	1.58	1.12	2.42	0.98	1.95
NUTDA-H	0.68	1.02	0.67	1.87	0.73	1.79

Unit of the tuning cost: number of 1000 n^3 operations.



(a) AMIS/MAMIS

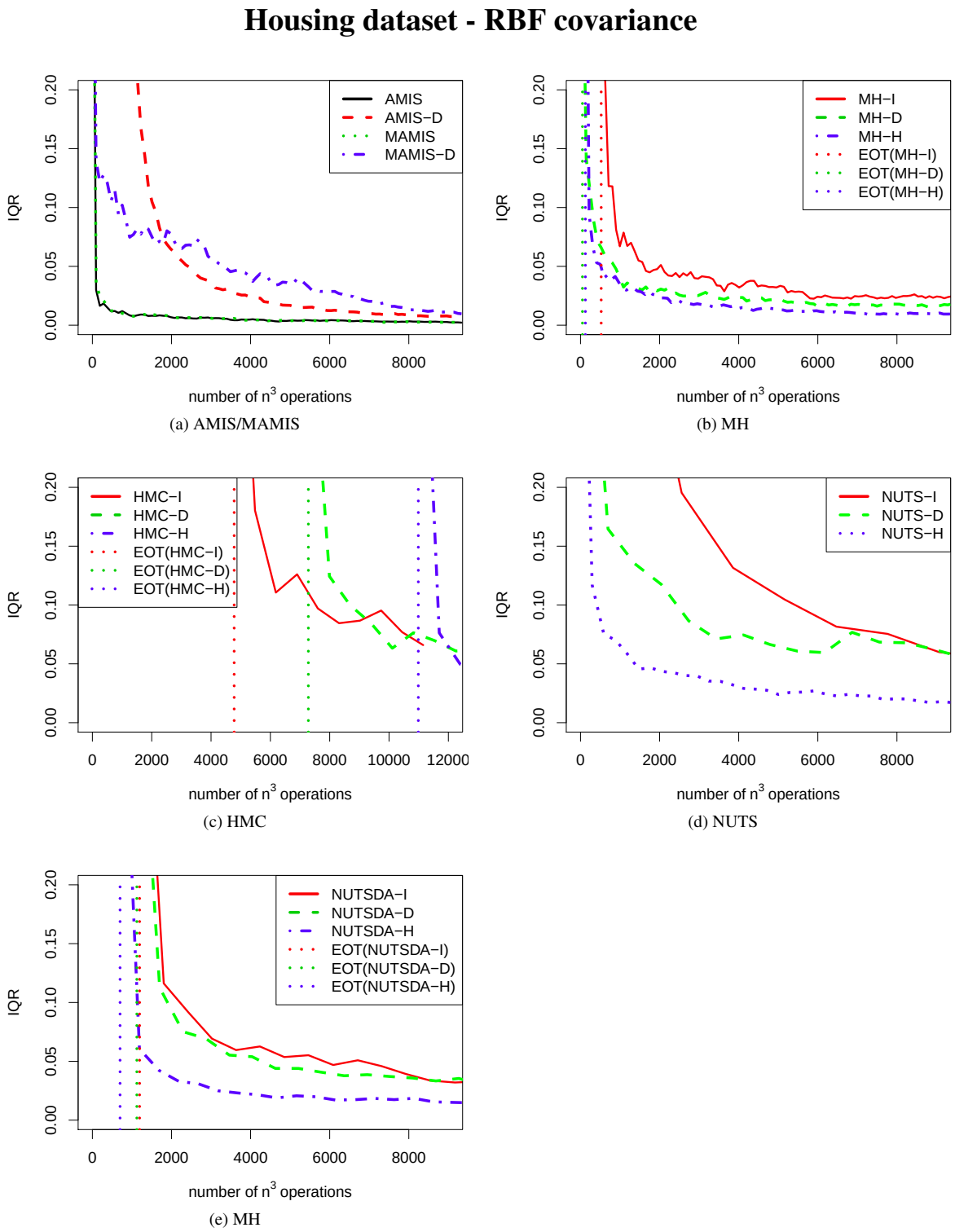
(b) MH

(c) HMC

(d) NUTS

(e) NUTSDA

Figure A.1 Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Concrete dataset (RBF covariance case). EOT stands for "end of tuning".



(a) AMIS/MAMIS

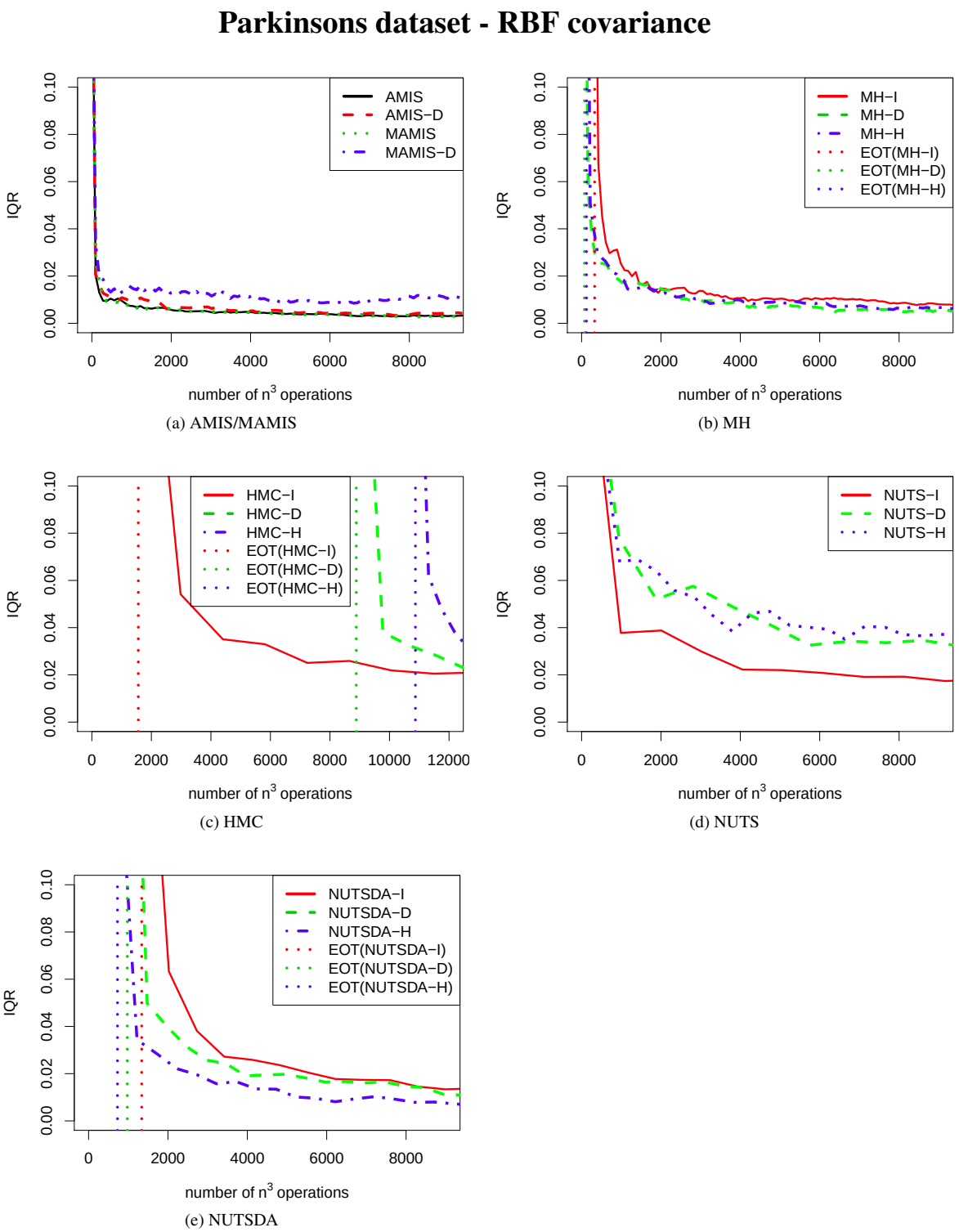
(b) MH

(c) HMC

(d) NUTS

(e) MH

Figure A.2 Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Housing dataset (RBF covariance case). EOT stands for "end of tuning".



(a) AMIS/MAMIS

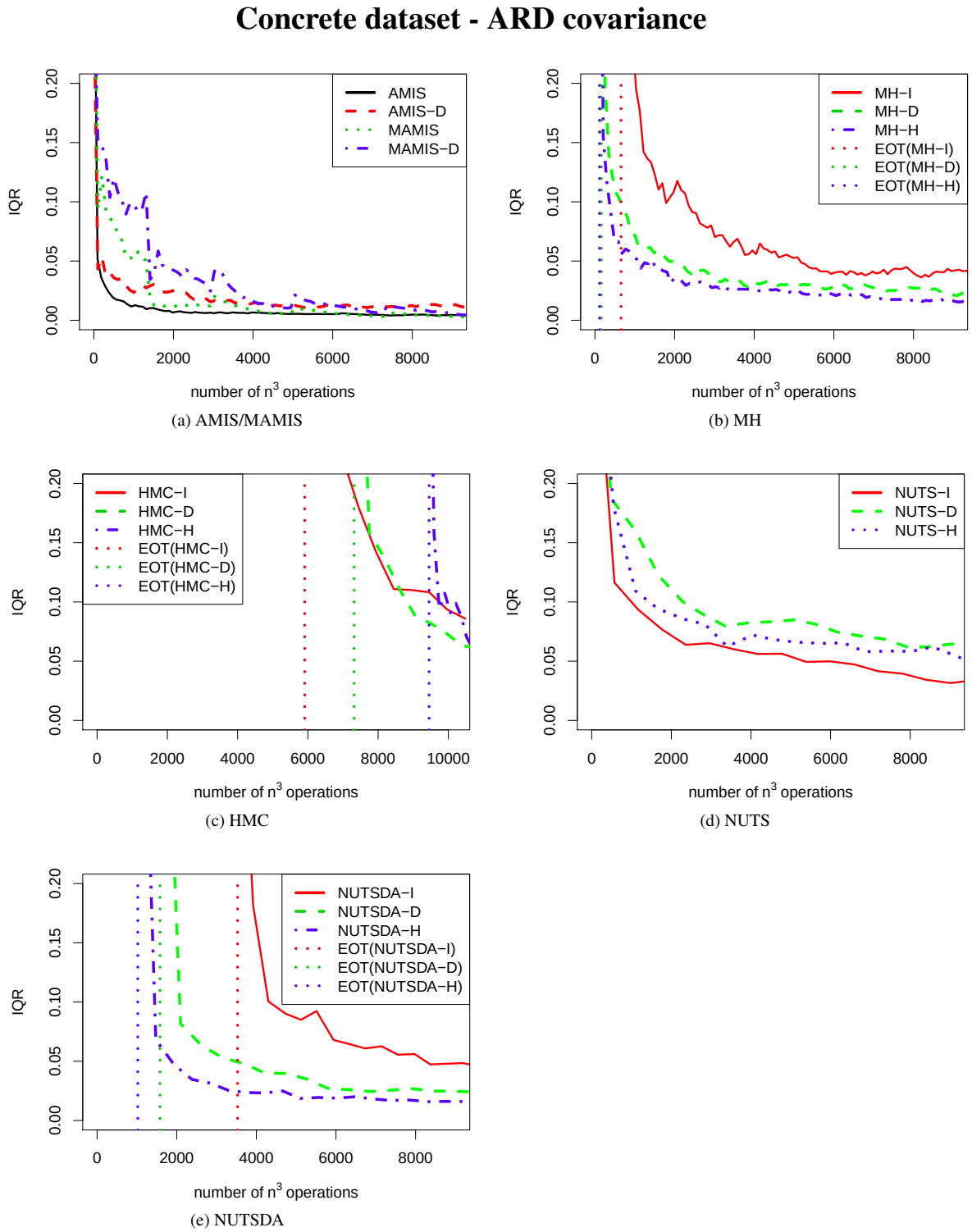
(b) MH

(c) HMC

(d) NUTS

(e) NUTSDA

Figure A.3 Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Parkinsons dataset (RBF covariance case). EOT stands for "end of tuning".



(a) AMIS/MAMIS

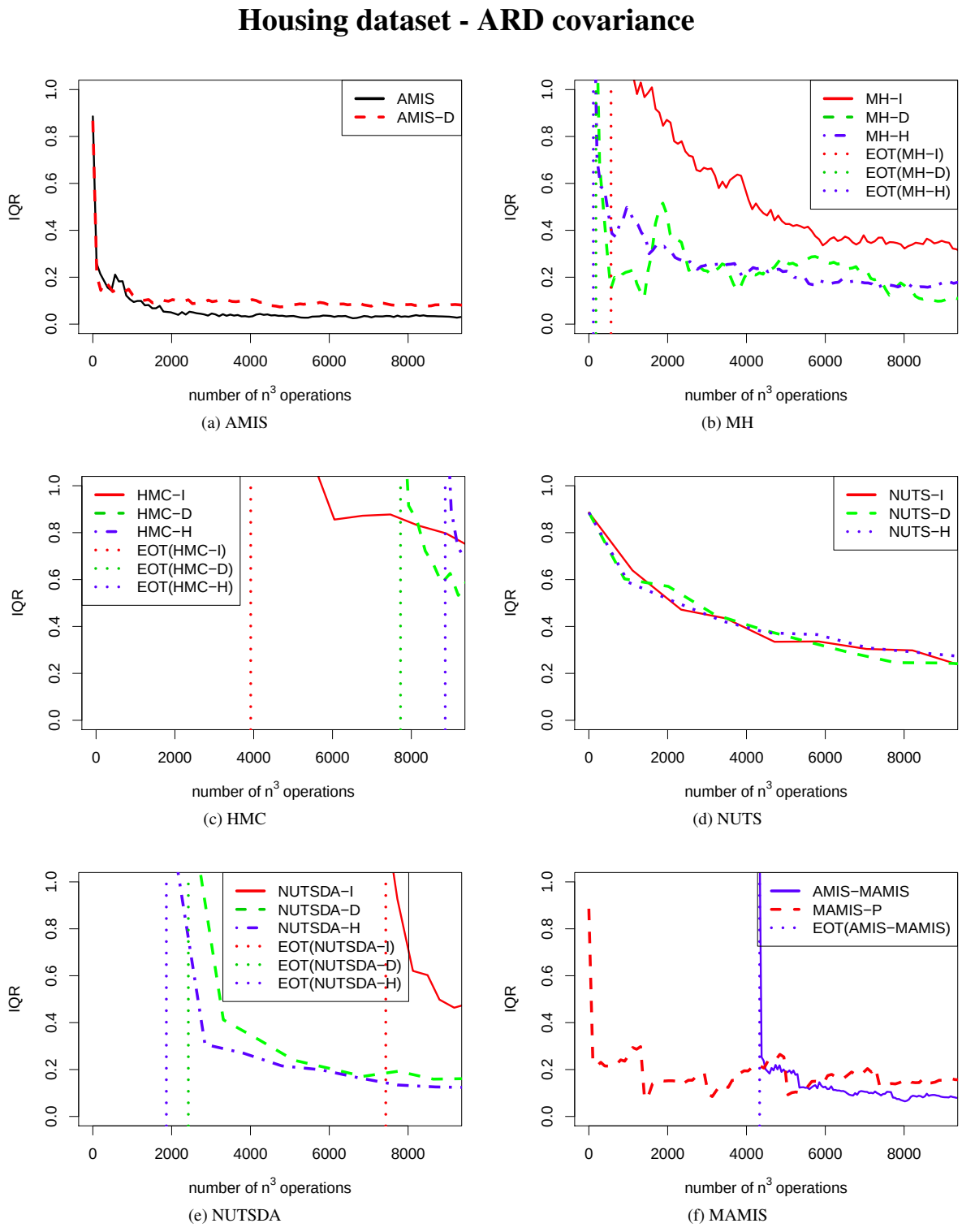
(b) MH

(c) HMC

(d) NUTS

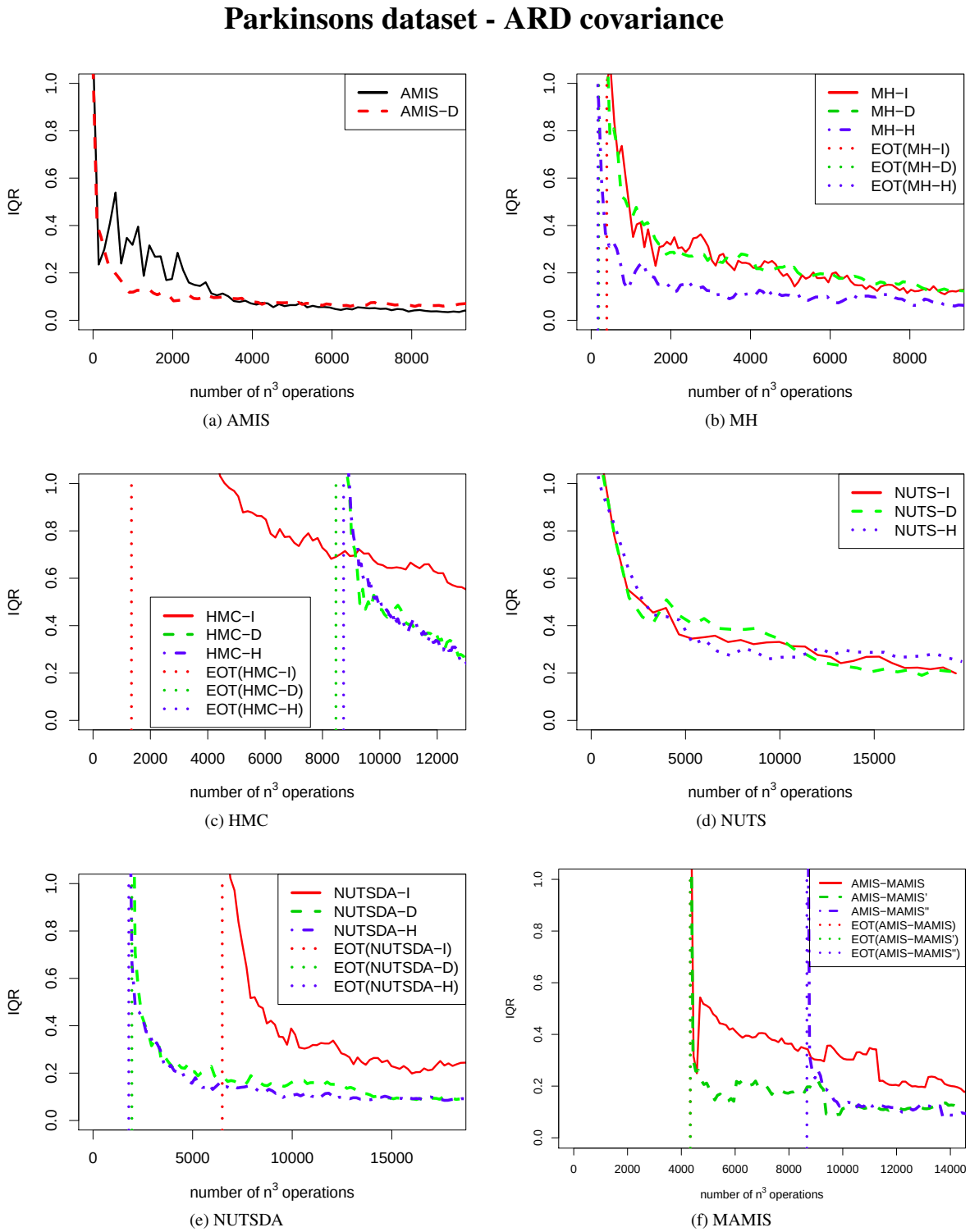
(e) NUTSDA

Figure A.4 Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Concrete dataset (ARD covariance case). EOT stands for "end of tuning".



(a) AMIS (b) MH (c) HMC (d) NUTS (e) NUTSDA (f) MAMIS

Figure A.5 Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Housing dataset (ARD covariance case). EOT stands for "end of tuning".



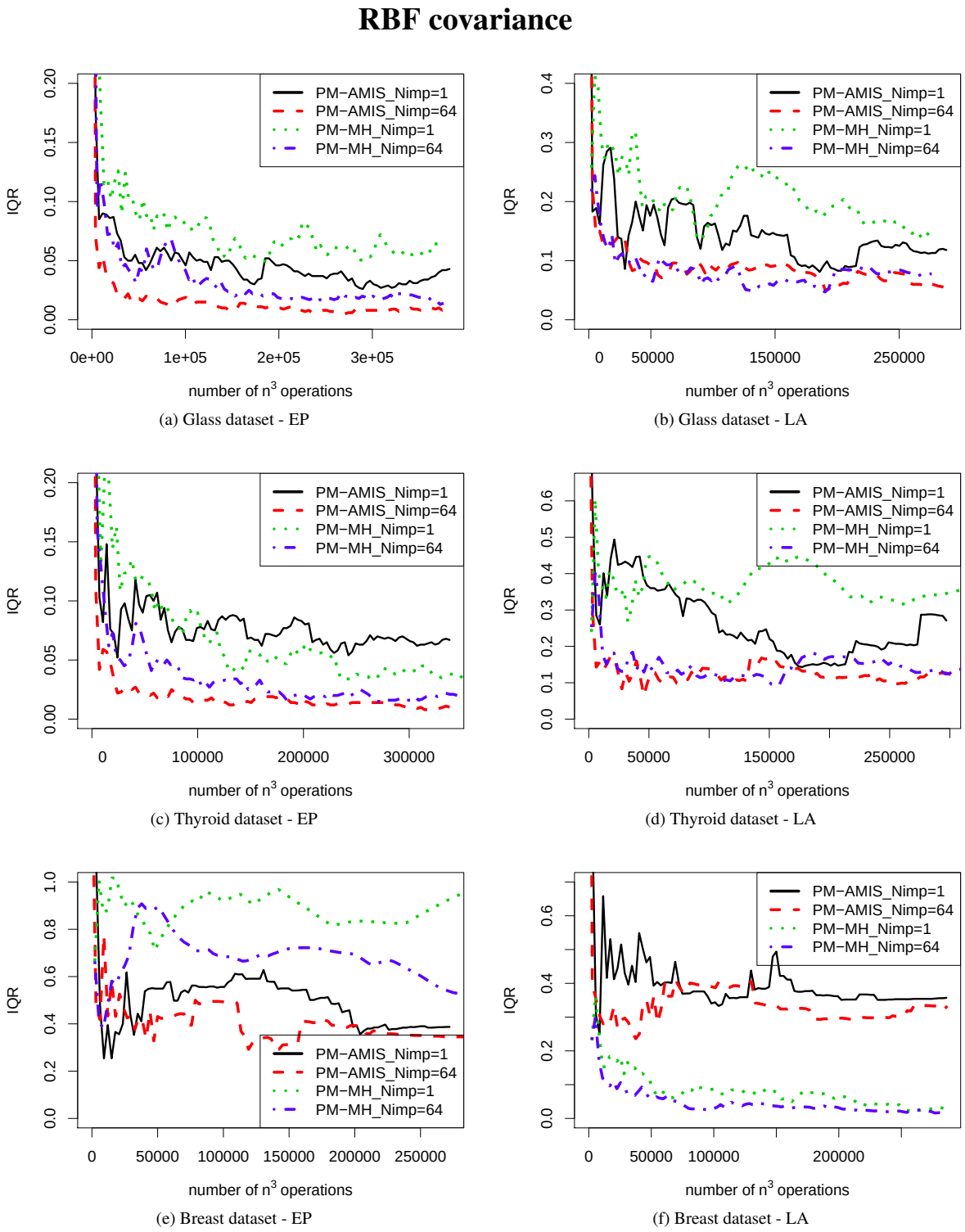
(a) AMIS (b) MH (c) HMC (d) NUTS (e) NUTSDA (f) MAMIS

Figure A.6 Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Parkinsons dataset (ARD covariance case). EOT stands for "end of tuning".

Appendix B

Convergence of samplers for GP classification

Appendix B shows the convergence results of PM-AMIS/PM-MH for the three classification data sets (Glass, Thyroid and Breast) for both the RBF (Figure B.1) and ARD (Figure B.2) covariances. In the figures, LA represents the case where the Gaussian approximation to the posterior of latent variables $\mathbf{f}$ is obtained by LA approximation, whereas EP denotes the case where the Gaussian approximation is obtained by EP approximation. Nimp denotes the number of importance samples of latent variables $\mathbf{f}$ to estimate the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. As can be seen from the figures, both the PM-AMIS and PM-MH algorithms with higher number of importance samples (Nimp=64) converge much faster than those with lower number of importance samples (Nimp=1) in both the EP and LA approximation cases as expected. The results also indicate that PM-AMIS is competitive with PM-MH in terms of convergence speed in most of the EP and LA approximation cases. Moreover, PM-AMIS/PM-MH seem to converge faster with EP approximation than with LA approximation in most cases which is probably because EP yields a more accurate approximation than LA as reported in [80, 104].



(a) Glass dataset - EP

(b) Glass dataset - LA

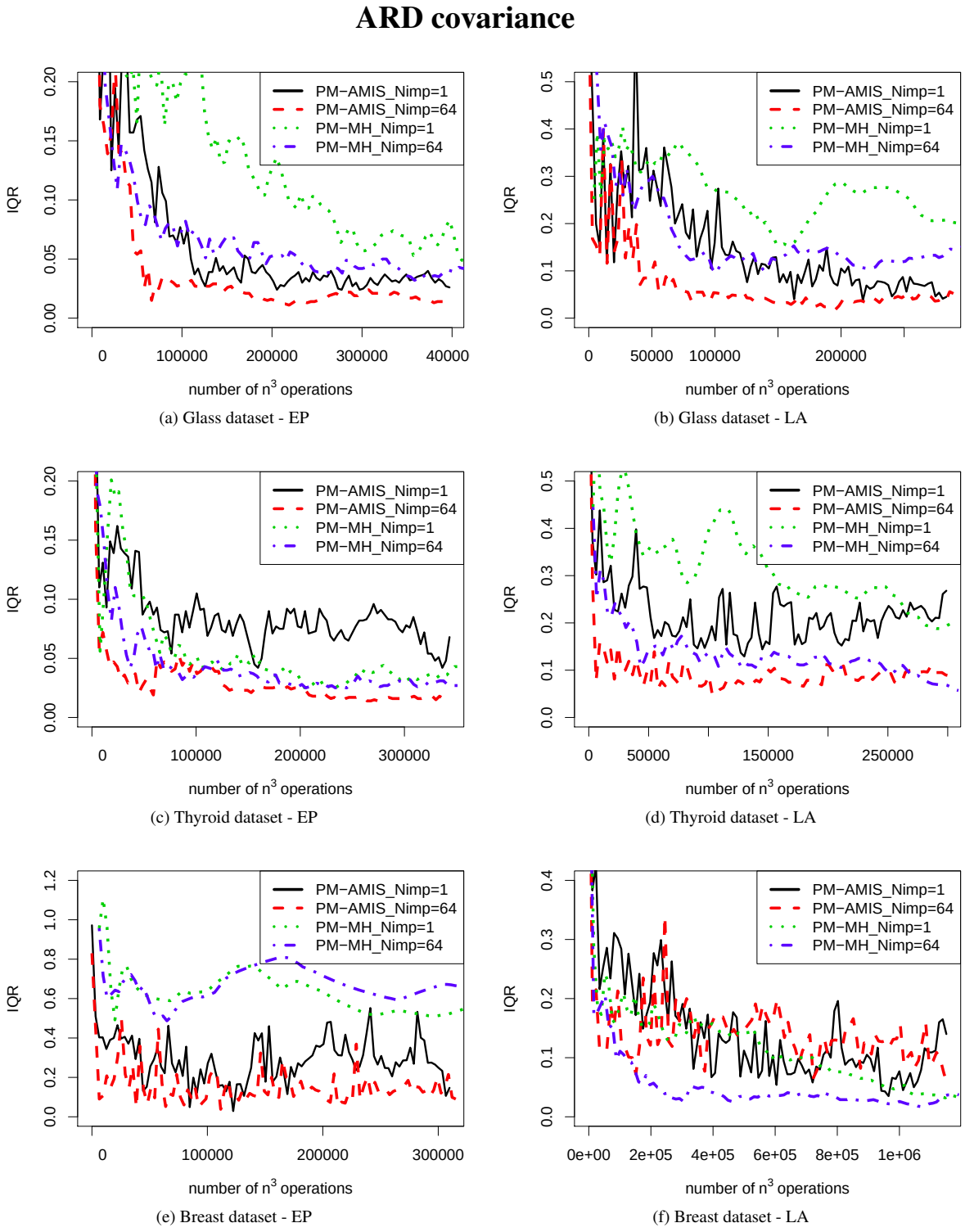
(c) Thyroid dataset - EP

(d) Thyroid dataset - LA

(e) Breast dataset - EP

(f) Breast dataset - LA

Figure B.1 Convergence of PM-AMIS, PM-MH for the RBF case. LA indicates the Gaussian approximation to the posterior of latent variables $\mathbf{f}$ is obtained by LA approximation, whereas EP indicates the Gaussian approximation is obtained by EP approximation. Nimp denotes the number of importance samples of latent variables to estimate the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$.



(a) Glass dataset - EP

(b) Glass dataset - LA

(c) Thyroid dataset - EP

(d) Thyroid dataset - LA

(e) Breast dataset - EP

(f) Breast dataset - LA

Figure B.2 Convergence of PM-AMIS, PM-MH for the ARD case. LA indicates the Gaussian approximation to the posterior of latent variables $\mathbf{f}$ is obtained by LA approximation, whereas EP indicates the Gaussian approximation is obtained by EP approximation. Nimp denotes the number of importance samples of latent variables to estimate the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$.

Appendix C

Mathematical Background

C.1 Gaussian Identities

The joint probability density of the multivariate Normal (or Gaussian) distribution is given by

$$p(\mathbf{x} \mid \mathbf{m}, \Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mathbf{m})\Big) \tag{C.1}$$

where $\mathbf{m}$ is the mean vector (of lenght D), and Σ is the (symmetric, positive definite) covariance matrix (of size $D \times D$). A shorthand for eq. (C.1) is

$$\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \Sigma) \tag{C.2}$$

Let $\mathbf{x}$, $\mathbf{y}$ be jointly Gaussian random vectors

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^{\mathrm{T}} & B \end{bmatrix} \right) \tag{C.3}$$

then the marginal distribution of $\mathbf{x}$ is

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, A) \tag{C.4}$$

and the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is

$$\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_y + C^{\mathrm{T}}A^{-1}(\mathbf{x} - \boldsymbol{\mu}_x), B - C^{\mathrm{T}}A^{-1}C) \tag{C.5}$$

The product of two Gaussians gives another unnormalised Gaussian

$$\begin{aligned} &\mathcal{N}(\mathbf{x} \mid \mathbf{a}, A)\mathcal{N}(\mathbf{x} \mid \mathbf{b}, B) = Z^{-1}\mathcal{N}(\mathbf{x} \mid \mathbf{c}, C) \quad \text{where} \\ &\mathbf{c} = C(A^{-1}\mathbf{a} + B^{-1}\mathbf{b}) \quad \text{and} \quad C = (A^{-1} + B^{-1})^{-1} \end{aligned} \tag{C.6}$$

and the normalising costant takes the Gaussian form (in $\mathbf{a}$ or $\mathbf{b}$)

$$Z^{-1} = (2\pi)^{-D/2}|A + B|^{-1/2} \exp\Big(-\frac{1}{2}(\mathbf{a} - \mathbf{b})^{\mathrm{T}}(A + B)^{-1}(\mathbf{a} - \mathbf{b})\Big) \tag{C.7}$$

Given a marginal Gaussian distribution of $\mathbf{x}$ and a conditional Gaussian distribution of $\mathbf{y}$ given $\mathbf{x}$ in the form:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Lambda^{-1}) \tag{C.8}$$

$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid A\mathbf{x} + \mathbf{b}, L^{-1}) \tag{C.9}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid A\boldsymbol{\mu} + \mathbf{b}, L^{-1} + A\Lambda^{-1}A^{\mathrm{T}}) \tag{C.10}$$

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\mathbf{x} \mid \Sigma\{A^{\mathbf{T}}L(\mathbf{y} - \mathbf{b}) + \Lambda\boldsymbol{\mu}\}, \Sigma) \tag{C.11}$$

where

$$\Sigma = (\Lambda + A^{\mathbf{T}}LA)^{-1} \tag{C.12}$$

C.2 Matrix Identities

Let Z be a $n \times n$ matrix, W be a $m \times m$ matrix, U and V be matrices both of size $n \times m$ and assume the relevant inverses all exist. Then the *matrix inversion lemma*, also known as the Woodbury, Sherman & Morrison formula is given by

$$(Z + UWV^{\mathrm{T}})^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^{\mathrm{T}}Z^{-1}U)^{-1}V^{\mathrm{T}}Z^{-1} \tag{C.13}$$

If Z^{-1} is known, a low rank (i.e. $m < n$) perturbation made to Z, as in left-hand side of eq. (C.13), will result in considerable speedup [121].

Bibliography

[1] C. Andrieu and C. P. Robert. Controlled MCMC for optimal sampling. *Bernoulli*, 9:395–422, 2001.

[2] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

[3] . Aristotle, E. Barker, and R. F. Stalley. *The Politics*. Oxford University Press, 1995.

[4] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.

[5] S. Bai, T. Zhu, and L. Cheng. Big-five personality prediction based on user behaviors at social network sites. Technical report, Cornell University, 2012.

[6] E. Bedrosian. The analytic signal representation of modulated waveforms. In *Prodeedings of the IRE*, volume 50, pages 2071 – 2076, 1962.

[7] E. Bedrosian. A product theorem for Hilbert transforms. In *Prodeedings of the IEEE*, volume 51, pages 868 – 869, 1963.

[8] B. Berlin. *Basic color terms: Their universality and evolution*. University of California Press, 1991.

[9] A. Beskos, N. Pillai, G. O. Roberts, J. M. Sanz-Serna, and A. M. Stuart. Optimal tuning of hybrid Monte Carlo algorithm. *Bernoulli*, 19:1501–1534, 2013.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[11] B. Boashas. Estimating and interpreting the instantaneous frequency of a signal - part 1: fundamentals. In *Prodeedings of the IEEE*, volume 80, pages 520 – 538, 1992.

[12] B. Boashash. *Time Frequency Signal Analysis and Processing*. Elsevier, 2003.

[13] J. W. Brown and R. V. Churchill. *Complex Variables and Applications*. McGraw-Hill Higher Education, 2009.

[14] O. Cappe, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13:907–929, 2004.

[15] F. Celli. Predicting personality with social media. In *Unsupervised personality recognition for social network sites*, pages 59–62, 2012.

[16] K. M. A. Chai. Variational multinomial logit Gaussian process. *Journal of Machine Learning Research*, 13(Jun):1745–1808, 2012.

[17] R. Chalfen. *Snapshot versions of life*. Bowling Green: Bowling Green State University Popular Press, 1987.

[18] W. Chu, Y. Chen, and K. Chen. Size does matter: How image size affects aesthetic perception? In *Proceedings of the ACM International Conference on Multimedia*, pages 53 – 62, 2013.

[19] S. Cloninger. Conceptual issues in personality theory. In P.J. Corr and G. Matthews, editors, *The Cambridge handbook of personality psychology*, pages 3–26. Cambridge University Press, 2009.

[20] L. Cohen. *Time-Frequency Analysis*. Prentice Hall, 1995.

[21] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603 – 619, 2002.

[22] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39:798–812, 2012.

[23] D. Coutu. We googled you. *Harvard Business Review*, 85(6):1–8, 2007.

[24] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 213–222, 2013.

[25] L. Csató and M. Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641 – 668, 2002.

[26] W. Curran, T. Moore, T. Kulesza, W. Wong, S. Todorovic, S. Stumpf, R. White, and M. Burnett. Towards recognizing cool: can end users help computer vision recognize subjective attributes of objects in images? In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 285–288. ACM, 2012.

[27] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer, 2006.

[28] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[29] R. Deering and J. F. Kaiser. The use of a masking signal to improve empirical mode decomposition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 485 – 488, 2005.

[30] R. Douc, A. Guillin, J. M. Marin, and C. Robert. Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 35:420–448, 2007a.

[31] R. Douc, A. Guillin, J. M. Marin, and C. Robert. Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007b.

[32] A. Doucet, N. D. Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. New York : Springer-Verlag, 2001.

[33] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[34] M. Duggan and L. Rainie. Cell phone activities 2012. Technical report, Pew Research Center, 2012.

[35] D. Evans, S. D. Gosling, and A. Carroll. What elements of an online social networking profile predict target-rater agreement in personality impressions. In *Proceedings of AAAI International Conference on Weblogs and Social Media*, pages 45 – 50, 2008.

[36] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. *Proceedings of the ACM International Conference on Multimedia*, pages 835–838, 2013.

[37] S. Farlow. *Partial differential equations for scientists and engineers*. Courier Corporation, 2012.

[38] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861 – 874, 2006.

[39] M. Filippone. Bayesian inference for Gaussian process classifiers with annealing and pseudo-marginal MCMC. In *22nd International Conference on Pattern Recognition*, pages 614–619, 2014.

[40] M. Filippone and M. Girolami. Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226, 2014.

[41] M. Filippone, A. F. Marquand, C. R. V. Blain, S. C. R. Williams, J. Mourão-Miranda, and M. Girolami. Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities. *The Annals of Applied Statistics*, 6(4):1883–1905, 2012.

[42] M. Filippone, M. Zhong, and M. Girolami. A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Machine Learning*, 93(1):93–114, 2013.

[43] S. Fitzgerald, D. Evans, and R. Green. Is your profile picture worth 1000 words? photo characteristics associated with personality impression agreement. In *Proceedings of AAAI International Conference on Weblogs and Social Media*, 2009.

[44] J. M. Flegal, M. Haran, and G. L. Jones. Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23(2):250–260, 2007.

[45] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach, 2nd Edition*. Pearson, 2012.

[46] Frauke Friedrichs and Christian Igel. Evolutionary tuning of multiple SVM parameters. *Neurocomputing*, 64:107–117, 2005.

[47] D. Funder. Personality. *Annual Reviews of Psychology*, 52:197 – 221, 2001.

[48] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers*, 93:429–441, 1946.

[49] D. Gabor. On an ambiguity in the definition of the amplitude and phase of a signal. *Signal Processing*, 79:301–307, 1999.

[50] A. Gelman, G. O. Roberts, and W. R. Gilks. Efficient Metropolis jumping rules. *Bayesian statistics*, 5(599-608), 1996.

[51] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[52] C. Georgescu. Synergism in low level vision. In *Proceedings of the International Conference on Pattern Recognition*, pages 150 – 155, 2002.

[53] W. Gilks, G. Roberts, and S. Sahu. Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, 93(443):1045–1054, 1998.

[54] Tobias Glasmachers and Christian Igel. Maximum likelihood model selection for 1-norm soft margin SVMs with multiple parameters. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1522–1528, 2010.

[55] J. Golbeck, C. Robles, and K. Turner. Predicting personality with social media. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems*, pages 253 – 262, 2011.

[56] D. Goleman. *Social Intelligence*. Hutchinson, 2006.

[57] R. Grossman, G. Seni, J. Elder, N. Agarwal, and H. Liu. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool, 2010.

[58] H. Haario, E. Saksman, and J. Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14:375–395, 1999.

[59] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

[60] H. Haario, E. Saksman, and J. Tamminen. Componentwise adaptation for MCMC. Technical Report Preprint 342, Dept. of Mathematics, University of Helsinki, 2003.

[61] R. Haralick and L. Shapiro. *Computer and Robot Vision*. Addison Wesley, 1992.

[62] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[63] J. Hensman, A. G. Matthews, M. Filippone, and Z. Ghahramani. MCMC for variationally sparse Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1648–1656, 2015.

[64] J. Hensman, A. G. D. G. Matthews, and Z. Ghahramani. Scalable Variational Gaussian Process Classification. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*, 2015.

[65] F. Hoenig. Defining computational aesthetics. In *Proceedings of the Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 13–18, 2005.

[66] M. D Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[67] N. A. V. House. Flickr and public image-sharing: distant closeness and photo exhibition. In *CHI 2007 Extended Abstracts on Human Factors in Computing Systems*, pages 2717 – 2722, 2007.

[68] N. A. V. House. Personal photography, digital technologies and the uses of the visual. *Visual Studies*, 26(2):125 – 134, 2011.

[69] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London A*, 1971,1998.

[70] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 145 – 152, 2011.

[71] M. Johnson. Subcortical face processing. *Nature Reviews Neuroscience*, 6(10):766 – 774, 2005.

[72] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[73] A. M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53:59–68, 2010.

[74] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[75] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli. Predictiong continuous conflilict perception with Bayesian Gaussian processes. *IEEE Transactions on Affective Computing*, 5(2):187–200, 2014.

[76] L. Kinsler, A. Frey, A. Coppens, and J. Sanders. *Fundamentals of Acoustics, 3rd Edition*. Wiley Publishing, 1982.

[77] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1137 – 1145, 1995.

[78] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. In *Proceeedings of the National Academy of Sciences*, volume 110, pages 5802 – 5805, 2013.

[79] R. Kulhavý. *Recursive nonlinear estimation: A geometric approach*. Springer, 1996.

[80] M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6(Oct):1679–1704, 2005.

[81] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5(Jan):27–72, 2004.

[82] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. Figueiras-Vidal. Sparse spectrum gaussian process regression. *Journal of Machine Learning Research*, 11(Jun):1865–1881, 2010.

[83] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. Springer-Verlag, 1994.

[84] M. Lipczak, M. Trevisiol, and A. Jaimes. Analyzing favorite behavior in flickr. In *International Conference on Multimedia Modeling*, pages 535–545. Springer, 2013.

[85] H. Liu. Social network profiles as taste performances. *Journal of Computer-Mediated Communication*, 13(1):252 – 275, 2007.

[86] S. Livens, P. Scheunders, G. V. D. Wouwer, and D. V. Dyck. Wavelets for texture analysis, an overview. In *Sixth International Conference on Image Processing and Its Applications*, volume 2, pages 581–585. IET, 1997.

[87] C. M. Lloyd, T. Gunter, M. A. Osborne, and S. J. Roberts. Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, pages 1814–1822, 2015.

[88] P. Lovato, A. Perina, N. Sebe, O. Zandoná, A. Montagnini, M. Bicego, and M. Cristani. Tell me what you like and I'll tell you what you are: discriminating visual preferences on Flickr data. In *Proceedings of the Asian Conference on Computer Vision*, 2012.

[89] J. Machajdik and A. Hanbury. Affective image classification using features inspired

by psychology and art theory. In *Proceedings of the ACM International Conference on Multimedia*, pages 83 – 92, 2010.

[90] D. J. C. MacKay. Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions*, 100(2):1053 –1062, 1994.

[91] K. Mardia and P. Jupp. *Directional Statistics*. Wiley, 2009.

[92] J. M. Marin, P. Pudlo, and M. Sedki. Consistency of the adaptive multiple importance sampling. *eprint arXiv:1211.2548v2*, 2014.

[93] A. G. D. G. Matthews, J. Hensman, R. E. Turner, and Z. Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*, 2016.

[94] G. Matthews, I. J. Deary, and M. C. Whiteman. *Personality traits*. Cambridge University Press, 2003.

[95] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[96] I. Murray and R. P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, pages 1732–1740, 2010.

[97] I. Murray and M. M. Graham. Pseudo-marginal slice sampling. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)*, 2016.

[98] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, September 1993.

[99] R. M. Neal. Regression and classification using Gaussian process priors (with discussion). *Bayesian Statistics*, 6:475–501, 1999.

[100] R. M. Neal. Slice Sampling. *The Annals of Statistics*, 31:705–767, 2003.

[101] R. M. Neal. *Handbook of Markov Monte Carlo, chapter 5: MCMC using Hamitonian dynamics*. CRC Press, 2011.

[102] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[103] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh. Towards discovery of influence and personality traits through social link prediction. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 566–569, 2011.

[104] H. Nickisch and C. E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, October 2008.

[105] A. H. Nuttall and E. Bedrosian. On the quadrature approximation to the Hilbert transform of modulated signals. In *Prodeedings of the IEEE*, volume 54, pages 1458 – 1459, 1966.

[106] M. S. Oh and J. O. Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168, 1992.

[107] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145 – 175, 2001.

[108] M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

[109] L. Ortiz and L. Kaelbling. Adaptive importance sampling for estimation in structured domains. In *Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 446–454, 2000.

[110] A. Owen and Y. Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

[111] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, 22(1):59–73, 2007.

[112] A. Pentland. Social signal processing. *IEEE Signal Processing Magazine*, 24(4):108–111, 2007.

[113] B. Picinbono. On instantaneous amplitude and phase of signals. *IEEE Transactions on Signal Processing*, 45(3):552 –560, 1997.

[114] M. K. Pitt, R. S. Silva, P. Giordani, and R. Kohn. On some properties of markov chain monte carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171:134–151, 2012.

[115] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[116] B. V. D. Pol. The nonlinear theory of electric oscillations. In *Prodeedings of the IRE*, volume 22, pages 1051 – 1086, 1934.

[117] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[118] L. Rainie, J. Brenner, and K. Purcell. Photos and videos as social currency online. Technical report, Pew Research Center, 2012.

[119] L. Rainie and B. Wellman. *Networked*. MIT Press, 2012.

[120] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007.

[121] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[122] R. Rato, M. D. Ortigueira, and A. Batista. On the HHT, its problems and some solutions. *Mechanical Systems and Signal Processing*, 22(6):1374 – 1394, 2008.

[123] S. Ray and D. Page. Multiple instance regression. In *Proceedings of the 18th International Conference on Machine Learning*, pages 425 – 432, 2001.

[124] C. J. V. Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.

[125] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

[126] G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2), 1997.

[127] G.O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

[128] D. Rubin. *Using the SIR algorithm to simulate posterior distributions. In Bayesian Statistics 3 (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 395-402.* Oxford Univ. Press., 1988.

[129] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[130] S. K. Sahu and A. A. Zhigljavsky. Self regenerative Markov chain Monte Carlo with adaptation. *Bernoulli*, 9(3), 2003.

[131] S. Sandoval and P. L. De Leon. Theory of the Hilbert spectrum. *arXiv preprint arXiv:1504.07554*, 2015.

[132] G. Saucier and L. R. Goldberg. The language of personality: Lexical perspectives. *The five-factor model of personality: Theoretical perspectives*, pages 21–50, 1996.

[133] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[134] C. Segalin, A. Perina, M. Cristani, and A. Vinciarelli. The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Transactions on Affective Computing*, 2016.

[135] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257 – 1264, 2005.

[136] A. Solin and S. Särkkä. Hilbert space methods for reduced-rank Gaussian process regression. *arXiv preprint arXiv:1401.5508*, 2014.

[137] J. Suler. Image, word, action: Interpersonal dynamics in a photo-sharing community. *CyberPsychology & Behavior*, 11(5):555 – 560, 2008.

[138] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460 – 473, 1978.

[139] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[140] B. M. Taylor and P. J. Diggle. INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284, 2014.

[141] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[142] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[143] M. K. Titsias. Variational learning of inducing variables in sparse Gaussian Processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 567–574, 2009.

[144] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin. A complete ensemble empirical mode decomposition with adaptive noise. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4144 – 4147, 2011.

[145] M. N. Tran, M. Scharth, M. K. Pitt, and R. Kohn. Importance sampling squared for Bayesian inference in latent variable models. *eprint arXiv:1309.3339*, 2014.

[146] D. E. Vakman. On the definition of concepts of amplitude phase and instantaneous frequency. *Radio Engineering and Electronic Physics*, 17:754–759, 1972.

[147] D. E. Vakman. Do we know what are the instantaneous frequency and instantaneous amplitude of a signal. *Radio Engineering and Electronic Physics*, 21:95–100, 1976.

[148] D. E. Vakman. Measuring the frequency of an analytical signal. *Radio Engineering and Electronic Physics*, 24:63–69, 1979.

[149] D. E. Vakman. On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *IEEE Transactions on Signal Processing*, 44(4):791–797, 1996.

[150] D. E. Vakman. *Signals, Oscillations and Waves*. Artech House, 1998.

[151] D. E. Vakman and L. A. Vainshtein. Amplitude, phase, frequency - fundamental concepts in the theory of oscillations. *Uspekhi Fizicheskikh Nauk*, 123:657 – 682, 1977.

[152] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.

[153] Michel F Valstar and Maja Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *International Workshop on Human-Computer Interaction*, pages 118–127. Springer, 2007.

[154] J. Vanhatalo and A. Vehtari. Sparse log Gaussian processes via MCMC for spatial epidemiology. *Journal of Machine Learning Research*, 1:73–89, 2007.

[155] S. Vazire and S. D. Gosling. e-perceptions: Personality impressions based on personal websites. *Journal of Personality and Social Psychology*, 87(1):123 –132, 2004.

[156] J. Ville. Theorie et applications de la notion de signal analytique. *Cables et Transmission*, 2a:61–74, 1948.

[157] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[158] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: survey of an emerging domain. *Image and Vision Computing*, 27:1743–1759, 2009.

[159] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder. Bridging the gap between social animal and unsocial machine: a survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):1743–1759, 2012.

[160] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511 – 518, 2001.

[161] V. Šmídl and R. Hofman. Efficient sequential Monte Carlo sampling for continuous monitoring of a radiation situation. *Technometrics*, 56(4):514–528, 2014.

[162] J. V. D. Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 – 8, 2007.

[163] B. Wellman and M. Gulia. The network basis of social support: A network is more than the sum of its ties. *Networks in the global village*, pages 83–118, 1999.

[164] C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1342–1351, 1998.

[165] A. Wright. Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, 5(3):292–296, 2014.

[166] Z. Wu and N. E. Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 1(01):1–41, 2009.

[167] X. G. Xia and L. Cohen. On analytic signals with nonnegative instantaneous frequency. In *Prodeedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[168] X. Xiong, M. Filippone, and A. Vinciarelli. Looking good with Flickr Faves: Gaussian processes for finding difference makers in personality impressions. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 412–415, 2016.

[169] X. Xiong, V. Šmídl, and M. Filippone. Adaptive multiple importance sampling for Gaussian processes. *Journal of Statistical Computation and Simulation*, 87(8):1644–1665, 2017.