Gaussian Processes for Spatiotemporal Modelling

Ricardo Andrade-Pacheco

Department of Computer Science

University of Sheffield

This dissertation is submitted for the degree of

Doctor of Philosophy

July 2015

To Buff

Acknowledgements

The completion of this PhD would not have been possible without the support of many people. First of all, I would like to thank Neil Lawrence for all his advice and constant encouragement. It has been a pleasure and a privilege being a collaborator of his group.

I also would like to thank my friends and colleagues in the Sheffield ML research group, for all their patience, their willingness to share their knowledge and for the good times we spent together. All my gratitude to Alan Saul, Alessandra Tosi, Alfredo Kalaitzis, Andreas Damianou, Arifur Rahman, Ciira Maina, Fariba Yousefi, James Hensman, Javier González, Jens Nielsen, Max Zwießele, Mike Smith, Mu Niu, Nicolas Durrande, Nicolo Fusi, Teo De Campos and Zhenwen Dai. I learnt a lot from each of them.

Thanks to John Quinn and Martin Mubangizi for always welcoming me with kindness and for their hard work in our collaboration.

Thanks to my parents, Carlos and Rosa, my brothers, Carlos and Alejandro, and to my extended family, Adriana Oropeza, Elizabeth Townend and Giovanni Torres; for all their support during these years.

Thanks to all the people who made of my time in Sheffield a wonderful experience.

This thesis was funded by CONACYT and SEP.

Abstract

A statistical framework for spatiotemporal modelling should ideally be able to assimilate different types of data from different of sources. Gaussian processes are commonly used tool for interpolating values across time and space domains. In this thesis we work on extending the Gaussian processes framework to deal with diverse noise model assumptions. We present a model based on a hybrid approach that combines some of the features of the discriminative and generative perspectives, allowing continuous dimensionality reduction of hybrid discrete-continuous data, discriminative classification with missing inputs and manifold learning informed by class labels.

We present an application of malaria density modelling across Uganda using administrative records. This disease represents a threat for approximately 3.3 billion people around the globe. The analysis of malaria based on the records available faces two main complications: noise induced by a highly variable rate of reporting health facilities; and lack of comparability across time, due to changes in districts delimitation. We define a Gaussian process model able to assimilate this features in the data and provide an insight on the generating process behind the records.

Finally, a method to monitor malaria case-counts is proposed. We use vector-valued covariance kernels to analyze the time series components individually. The short term variations of the infection are divided into four cyclical phases. The graphical tool provided can help quick response planning and resources allocation.

Table of contents

List of figures

List of tables

Symbols and Notation

Expectation propagation nomenclature

ℓ_i	Likelihood approximation factors with moment parameters $\tilde{\mu}_i, \tilde{\sigma}_i^2, \tilde{Z}_i$
$\mathbf{L}$	Cholesky decomposition of $\mathbf{K_{uu}} + \mathbf{K_{uf}}\tilde{\mathbf{\Sigma}}^{-1}\mathbf{K_{fu}}$
$\hat{\mathbf{L}}$	A diagonal matrix
q_{-i}	Cavity distribution with moment parameters μ_{-i}, σ^2_{-i}
$\mathbf{R}$	Cholesky decomposition of $\mathbf{K}_{\mathbf{uu}}^{-1}$
$\mathbf{s}_i$	Column i of the matrix $\mathbf{\Sigma}$
Z	Normalizing constant
${}_uA$	Update computation of a quantity A in an iterative algorithm
$\hat{\mathbf{\Psi}}$	Matrix in $\mathbb{R}^{n\times m}$
$\hat{\boldsymbol{\psi}}_i$	Column i of $\hat{\mathbf{\Psi}}$
$\boldsymbol{\gamma}$	Vector in $\mathbb{R}^m$
$\boldsymbol{\mu}$	Mean of the EP posterior approximation
$\boldsymbol{\omega}$	Vector in $\mathbb{R}^n$
$\mathbf{\Sigma}$	Variance of the EP posterior approximation
$\boldsymbol{\theta}$	Array of natural parameters of a Gaussian distribution $\boldsymbol{\theta} = (\tau, \nu)^\top$
τ	One of the natural parameters of a Gaussian distribution ($\tau = \sigma^{-2}$)
ν	One of the natural parameters of a Gaussian distribution ($\nu = \tau\mu$)

Functions

$\text{cov}(\cdot, \cdot)$	Covariance function
$g(\cdot)$	Monotonic continuous function
$\mathcal{I}_{\{x \in A\}}$	Index function with value 1 if $x \in A$ and zero otherwise
$K(\cdot, \cdot)$	Scalar-valued kernel function
$\mathcal{M}(\cdot)$	Mean function
$\text{var}(\cdot)$	Variance function
$\lambda(\cdot)$	Intensity function
$\mu(\cdot)$	Intensity measure
$\sigma(\cdot)$	Sigmoidal function
$\Gamma(\cdot, \cdot)$	Vector-valued kernel function

Model families

$\mathcal{GP}(\mathcal{M}, K)$	Gaussian process with mean function $\mathcal{M}$ and covariance function K
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{PP}(\lambda)$	Poisson process with intensity function λ

HMIS model parameters

r_i	Number of health facility units reporting in the i-th observation
t_i	Time of the i-th observation
δ_i	Error term with heterogeneous variance of the i-th observation
$\boldsymbol{\gamma}$	Nested-mean kernel components
$\hat{\omega}_i$	NDVI interpolated value
$\boldsymbol{\tau}$	Mean bias in district counts
ϵ_i	Error term with homogeneous variance of the i-th observation
ζ_i	Error term in the reporting process observed in the i-th observation

Inputs and outputs

$\mathbf{X}$	Matrix of inputs
$\mathbf{x}_*$	Test input
$\mathbf{x}_i$	The i-th training input
$\mathbf{Y}$	Matrix of outputs
y_*	Test output
$\mathbf{y}$	Vector of training outputs
y_i	The i-th training output
$\mathbf{Z}$	Matrix of inducing inputs
$\mathbf{z}_i$	The i-th inducing input

Superindices

d	Number of outputs or tasks
p	Output dimensionality
q	Input dimensionality

Latent variables

$\mathbf{f}_*$	Vector of latent variables at new input locations
$\mathbf{f}$	Vector of latent variables
$f_{\mathbf{x}_i}$	Scalar-valued latent function evaluated at $\mathbf{x}_i$
$h_{\mathbf{x}_i}$	Vector-valued latent function evaluated at $\mathbf{x}_i$
$\mathbf{u}$	Vector of inducing latent variables
$u_{\mathbf{z}_i}$	Inducing latent function (scalar-valued) evaluated at $\mathbf{x}_i$

Covariance terms

$\mathbf{B}$	Coregionalization matrix
$\mathbf{D}_{**}$	Diagonal of matrix $\mathbf{K}_{**} - \mathbf{Q}_{**}$

$\mathbf{D_{ff}}$	Diagonal of matrix $\mathbf{K_{ff}} - \mathbf{Q_{ff}}$
$\mathbf{K_{**}}$	Covariance matrix $K(\mathbf{X}_*, \mathbf{X}_*)$
$\mathbf{k_{f*}}$	Cross-covariance vector $K(\mathbf{X}, \mathbf{x}_*)$
$\mathbf{K_{ff}}$	Covariance matrix $K(\mathbf{X}, \mathbf{X})$
$\mathbf{K_{fu}}$	Covariance matrix $K(\mathbf{X}, \mathbf{Z})$
$\mathbf{K_{uu}}$	Covariance matrix $K(\mathbf{Z}, \mathbf{Z})$
k_{**}	Cross-covariance term $K(\mathbf{x}_*, \mathbf{x}_*)$
$\mathbf{Q_{**}}$	$\mathbf{K_{*u}K_{uu}^{-1}K_{u*}}$
$\mathbf{Q_{f*}}$	$\mathbf{K_{fu}K_{uu}^{-1}K_{u*}}$
$\mathbf{Q_{ff}}$	$\mathbf{K_{fu}K_{uu}^{-1}K_{uf}}$

Other symbols

$\mathrm{chol}(\cdot)$	Cholesky decomposition of a matrix
$\mathrm{diag}\,(\cdot)$	Vector with the diagonal elements of a matrix
$\mathbf{e}_i$	Basis vector with i-th entry equal to 1 and zero everywhere else
$\mathbf{I}$	Identity matrix
$\mathrm{KL}\,(\cdot\|\|\cdot)$	Kullback-Leibler divergence between two distributions
$\mathcal{L}$	Log-likelihood lower bound
m	Number of inducing inputs
n	Number of training cases
$\mathring{n}$	Cumulative counting process
$\mathcal{O}(\cdot)$	Execution time or space used
X	Random variable
Y	Random variable
p	Probability distribution

q	Probability distribution
Z	Random variable
$(f_s)_{s\in\mathbb{S}}$	Stochastic process with input space $\mathbb{S}$
$(u_s)_{s\in\mathbb{S}}$	Stochastic process with input space $\mathbb{S}$
η	Linear predictor in a generalized linear model
$\mathbf{\Psi}_1$	$\langle \mathbf{K_{uf}} \rangle_{q(\mathbf{X})}$
$\tilde{\mathbf{\Psi}}_2$	$\langle \mathbf{K_{uf}} \tilde{\mathbf{\Sigma}}^{-1} \mathbf{K_{fu}} \rangle_{q(\mathbf{X})}$
$\tilde{\psi}_0$	$\mathrm{tr}\left(\tilde{\mathbf{\Sigma}}^{-1} \langle \mathbf{K_{ff}} \rangle_{q(\mathbf{X})} \right)$
$\jmath(\cdot)$	Output or task index of a data instance
$< \cdot, \cdot >$	Inner product
$\langle \cdot \rangle_p$	Expectation under distribution p (when not relevant, p is omitted)
$[a_{ij}]$	Matrix with entries a_{ij}

General model parameters

ℓ	Kernel lengthscale
$\mathbf{w}$	Parameter vector with real entries
α	Scale parameter of a linear kernel
$\boldsymbol{\beta}$	Parameter vector with real entries
ϵ	Noise term
$\boldsymbol{\kappa}$	Parameter vector with real entries
ϵ	Mean of a Gaussian distribution
$\boldsymbol{\mu}$	First moment parameter of a multivariate Gaussian distribution
ρ	Correlation between outputs
$\mathbf{\Sigma}$	Second central moment parameter of a multivariate Gaussian distribution
σ^2	Variance or scale parameter

Spaces

$\mathcal{B}$	Space of bounded linear operators
$\mathcal{H}$	Hilbert space
$\mathcal{Y}$	Euclidean or Hilbert space

Sets

$\mathbb{B}$	Bounded set
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{S}$	Non-empty set
$\mathbb{Z}$	The set of integer numbers
$\mathcal{J}$	Index set of output (task) components $\{1, \ldots, d\}$
$\mathcal{X}$	Subset of inputs such that $\mathcal{X} \in \{\mathbf{Z}, \mathbf{X}, \mathbf{X}_*\}$

Acronyms

AR	Autoregressive
ARCH	Autoregressive conditional heteroscedastic
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
DTC	Deterministic training conditional approximation
EP	Expectation propagation
EP-DTC	Sparse EP algorithm based on the DTC approximation
EP-FITC	Sparse EP algorithm based on the FITC approximation
FITC	Fully independent training conditional approximation
GARCH	Generalized autoregressive conditional heteroscedastic
GLM	Generalized linear model

GP	Gaussian process
GP-LVM	Gaussian process latent variable model
HIS	Health information system
HMIS	Health management information system
ICM	Intrinsic coregionalization model
KL	Kullback-Leibler
l.h.s.	Left hand side (of an equation)
LCM	Linear coregionalization model
LOO-CV	Leave-one-out cross-validation
MA	Moving average
MCMC	Markov chain Monte Carlo
NDVI	Normalized difference vegetation index
NIR	Near infrared light reflected
PCA	Principal components analysis
r.h.s.	Right hand side (of an equation)
RBF	Radial basis function or squared exponentiated kernel
RKHS	Reproducing kernel Hilbert space
VAR	Vector autoregressive
Var EP-DTC	Sparse EP algorithm used on the sparse variational approximation
VARIMA	Vector autoregressive integrated moving average
VARMA	Vector autoregressive moving average
VIR	Visible radiation
w.r.t.	With respect to
WHO	World health organization

Chapter 1

Time, Space and Uncertainty

Imagine a couple of dancers whose movements blend with the background music in a bar. The synchrony of the couple's motion with the rhythm of the music allows the spectators to anticipate their actions across the dancing floor. The immediate future of the couple's performance seems to be declared beforehand, in the same way a rolling ball communicates its new direction at every moment[1]. This is the essence of spatiotemporal modelling, where *patterns* are sought in the past and across space to understand the present better at different places and, maybe, have a glimpse into the future.

But patterns are a vague idea, they are just a structure in the perception of the observer. They are not really out there in the world, they are an imprint in the subjective experience of an individual. Yet, patterns are an echo of the world that sometimes carries a meaning. They can be the key of an answer to a *why?* Or to a *how?* The *perceptual organization* is part of our learning engine that helps turning plain data into knowledge (Wagemans et al., 2012). The whole and the sum of the parts are two different objects we learn from. We try to understand the part-features by studying the whole-features as much as the other way around. In this regard, the field of *statistics* looks for comprehensive structures that describe the data (Cressie and Wikle, 2011). It provides a principled learning mechanism that minimizes the subjectivity of the perceptual experience. The field of *machine learning*, on the other hand, is devoted to provide automated methods for pattern recognition and data analysis (Murphy, 2012). This means to endow a machine with the skill of perceptual organization, so that it can emulate our learning process.

[1]This example, perfect for motivating a discussion about space and time modelling, was originally used by Bergson (1889) in his doctoral thesis, translated to English as *Time and Free Will: An Essay on the Immediate Data of Consciousness.*

Statistics and machine learning share a common ground in which both attempt to provide answers with respect to our surrounding. Both build models as a tool for understanding a world where uncertainty is ubiquitous, starting for the data we collect. Our imperfect knowledge of reality, the reason for the world to look to us as random and sometimes even chaotic, has found a place within these models in the form of stochastic components (Chilès and Delfiner, 2009). These models have evolved, and keep evolving, to reflect our better understanding of the world, as well as our better understanding of our own epistemic uncertainty. The story of this evolution is interesting by itself, for it is a *memoire* of how scientists from distant places have contributed, across generations, to a common goal: tell how and, if possible, tell why.

1.1 Notes on the Evolution of Time Series Models

The structure dependency of a time series usually identified through patterns such as *trends*[2], *cyclic effects*[3] or *irregular fluctuations*[4]. A not uncommon approach for time series analysis is to decompose the observed variation of the series into signals that represent these patterns (Baxter and King, 1999; Cleveland and Tiao, 1976; Hyvärinen and Oja, 2000). Early models assumed that time series were either deterministic or at most disturbed by a single stochastic element, which accounted for the residual variation and had no significance in the structure of the series (Schuster, 1898). The idea of stochastic time processes with a more complex dependence structure was pioneered by Yule (1927) and Slutzky (1927), in their formulation of the *autoregressive* (AR) and *moving average* (MA) models. Both authors assumed a time series to be generated by uncorrelated random shocks with zero mean and constant variance[5].

Since an early stage, most of the time series literature has been developed within the scope the stationary theory of statistics. Given the difficulty of ensuring identical conditions for an ensemble of observations across time, stationarity becomes desirable as it provides a theoretical framework where all instants in a time series are practically equivalent[6]. The link between stationarity and stochastic time models was pointed out

[2]A trend represents the observed long term change in the mean of the series.

[3]Cyclic effects are variations around the trend. When these variations are regular and occur in annual periods, they are usually called seasonal effect.

[4]This fluctuations are a residual variation that do not correspond to the previous cases and might or might not be random.

[5]AR models assume that a stochastic process can be explained as a linear combination of previous realizations of the process plus a random shock. Meanwhile, MA models assume that the current observation of the process can be obtained by regressing over past random shocks.

[6] A strictly stationary process is characterized for having a distribution invariant under arbitrary time shifts. A less restrictive kind of stationarity, known as weak or second order stationarity, only

by Wold (1938)[7]. From Wold's decomposition theorem, it follows that AR processes can also be expressed as MA processes. The duality between AR and MA processes can be exploited by combining both into a single one, know as *autoregressive moving average* (ARMA) model. The advantage being that the resulting model can sometimes describe a time series with fewer parameters than the ones needed by a single AR or MA (Chatfield, 2013).

The general theory of *reproducing kernel Hilbert spaces* (RKHS) arose in the early 1940s (Aronszajn, 1943), and soon was started to be used by probabilists to study the structure of time series (Karhunen, 1947; Loève, 1948). Later, Parzen (1959, 1961, 1970) showed how RKHS is a natural setting for solving inference problems in time series.

Before the seventies, model selection was strongly dependent on expert criteria, as no algorithm was able to specify a model uniquely (De Gooijer and Hyndman, 2006). Box and Jenkins (1976) presented a principled and unifying framework, which allowed model identification, parameter estimation and diagnostic checking. Although they mainly focused on discrete time series with evenly spaced observations, their work contributed to the widespread use of time series techniques. Their framework can also be applied to some non-stationary time series by considering, as Yaglom (1955) did, processes whose differences are stationary. This way, the *autoregressive integrated moving average*[8] (ARIMA) model allows working with series with changing means due to trends or seasonal patterns[9]. Impulse-response models for *open loop systems*[10], can be defined by incorporating a *transfer function* in the ARIMA framework. For dealing with systems where variables interrelation defines a *closed loop*[11], a generalization to the multivariate case of the ARIMA framework can also be defined following Quenouille

requires a process to have a mean and covariance functions that do not depend on time shifts (Grigoriu, 2002).

[7]Wold's decomposition theorem states that a zero mean and second order stationary process can be decomposed as a deterministic time series plus a weighted sum of random uncorrelated time series.

[8]Broadly speaking, a stationary series is generated after differentiating, up to some order k, a non-stationary series and then applying an ARMA model to the new series. The term *integrated* means that the new series has to be added up to represent the original non-stationary series.

[9]More specifically, series with cyclic patterns require a *seasonal ARIMA* model. Although this is a more complex model than ARIMA, it is built using the same principles.

[10]Systems where a variable X has an effect on a variable Y, but not vice versa. In these systems X is regarded as input and Y is regarded as output.

[11]Two variables Y and X are in a closed loop when they affect each other. This is as opposed to an *open loop* or impulse-response system, where an input X has an effect on an output Y, but there is no feedback from Y to X. For closed loop systems the terms input and output are not appropriate anymore.

(1957). In practice, this generalization consists of re-expressing the ARIMA-type models in vector notation[12], leading to models known as *VAR*, *VARMA* or *VARIMA*.

Different extensions were proposed in the forthcoming years. For example, extensions that deal not only with non-stationarity due to the mean, but also due to a changing variance (Bollerslev, 1986; Engle, 1982)[13]. This is a fruitful field that has kept evolving since its very beginning. There are still problems to be solved and areas where methodologies available can be improved. One particularly interesting is an efficient integration of space and time modelling able to deal the computational limitations when using large data sets (Särkkä et al., 2013).

1.2 Notes on the Evolution of Geostatistics

There was a time when geography, following a regional approach, was mostly focused on describing and inventorying the characteristics of a place. A theoretical framework that included statistical inference was yet to arise during the *quantitative revolution*[14] in the 1950s and 1960s (Barnes, 2001; Burton, 1963). The statistical methods known by the geographer's of the first half of the XX century, were usually preceded by the mantra *independent and identically distributed*, and therefore were of little use in a field where spatial closeness tends to be opposed to independence and data are rarely obtained under identical conditions. Although spatial dependence was not disregarded by the statisticians of the time (Fisher, 1935), the development of a framework analogue to time series, which was already in progress, was constrained by the theoretical differences, and the mathematical complexities involved, between dependence in time and dependence in space (Whittle, 1954).

Matheron (1962, 1963) and Gandin (1963), independently, were the first to develop a *best linear unbiased predictor* for spatial modelling (Cressie, 1990), in terms of the optimal prediction theory developed by Wold (1938), Kolmogorov (1941) and Wiener (1942)[15]. The use of *Kriging*, as Matheron defined the predictor, became the characteristic feature of a new field known as *geostatistics*, concerned with continuous spatial variation.

[12]The interpretations of this generalization has a deeper meaning than just a re-expression. This topic will be further discussed later.

[13]The topic of *heteroscedasticity* will be discussed in this thesis, but not under the same approach of ARCH and GARCH models.

[14]The quantitative revolution is identified as a period of rapid transformation of the discipline due to its mathematization.

[15]Masani (1966) presents a brief, but detailed narration of how these authors contributed to prediction theory.

Within spatial statistics literature, geostatistical methods have traditionally been related to data observed at a set of spatial locations indexed in a continuous space[16]. This is as opposed to *lattice methods* and *point pattern methods*. The former related to data observed on a fixed set locations (not necessarily a regular grid) and the later used for data associated to a point process[17]. Lattices that represent spatial data aggregated by regions are often modelled as Markov Random Fields (Besag et al., 1991). A few examples of models to handle point patterns are presented in Appendix C.

Diggle et al. (2013) take a different approach and consider that distinctions based on data formats are not always appropriate. They argue that the main theoretical distinction within spatial statistics is the continuity or non-continuity of the process being modelled. The *model-based* approach (Diggle et al., 1998) has become the current paradigm for modelling variability and quantifying uncertainty: a hierarchical thinking that explicitly assumes a stochastic model, but also acknowledges a different uncertainty in the data and in the parameters of the process (Cressie and Wikle, 2011).

1.3 Structure Dependence and RKHS

The structure dependence is a cornerstone in stochastic modelling. While time series focuses on a dependence that is unidimensional and unidirectional, geostatistics deals with a multidimensional phenomenon that occurs in every direction. Despite these differences, the concepts of correlation or covariance are robust enough to provide a dependence model for both cases. Both Kriging and ARIMA models generate an interpolation function based on a covariance (or variogram) model derived from the data. It is often the case, in these areas, that correlation functions are not used to describe the association between different variables, as in the approach of Galton (1886) and Pearson (1920), but to describe the similarity of the values taken by the same variable across a domain. Hence the term *autocorrelation* is usually preferred[18].

RKHS draw a correspondence between a positive kernel and a *Hilbert space* of functions, through a series of so called *representation theorems*. This space, where the closeness of a function f to a function g means that the values of $f(x)$ are close to the values of $g(x)$, provides enough tools for doing inference with stochastic processes.

[16]In other words, if y_s is a data point observed at location s, then $s \in \mathbb{R}^q$, for some $q \in \mathbb{N}$.

[17]Point patters are used when the question of interest is the location of the events, rather than the intensity of a function across space.

[18]The term *autocorrelation* has been part of the standard jargon in the time series literature for long (see the translated work of Slutzky (1927), for example). In geostatistics, concepts like *variogram* or *correlogram* are commonly used instead (Cressie, 1992). The term *spatial autocorrelation* was first introduced in the late 1960s by Cliff and Ord (1969).

A key point in the intersection between RKHS and statistical theory is a theorem by Loève (1948) that links the class of positive functions to the class of covariance functions. This theorem opens the door for translating stochastic problems into functional ones (Berlinet and Thomas-Agnan, 2004). Following this path, Parzen (1959) exploited Mercer and Karhunen representation theorems (Karhunen, 1947; Riesz and Sz-Nagy, 1955) to define formal solutions to best linear prediction problems for stochastic processes.

Based on the bijection defined by RKHS, kernel functions enable to analyze non-linear patterns by embedding an inference problem into an abstract space with a *convenient structure*, so that it can be linearized. But this is not the only advantage of kernel-based learning methods. They also allow working with high-dimensional data at a low computational cost[19] without compromising the representation power (Shawe-Taylor and Cristianini, 2004).

1.4 The Gaussian Measure

When it comes to studying uncertainty, the Gaussian distribution is one of the most used distributions. Simplicity in its definition and convenience of its analytical properties can, of course, be on the list of reasons for assuming a process to be Gaussian. But the reason behind its widespread use goes beyond the fact that it is a relatively easy distribution to use. Gaussian distributions tend to arise naturally in some physically meaningful mathematical situations (Sudakov, 1993). As an example, consider the *central limit theorem* and its extensions to the infinite dimensional case.

Among the nice properties of the Gaussian distributions, they are completely defined by their first two moments. Moreover, centered Gaussian distributions are uniquely defined by their covariance. As Abrahamsen (1997) puts it, the study of Gaussian processes (GP) is in many ways the study of covariance functions. But if a covariance function fully characterizes a Gaussian process, then a kernel does it as well[20]. Thus, there is even a bijection between RKHS and Gaussian processes (Berlinet and Thomas-Agnan, 2004; Hein and Bousquet, 2004). As a result, the use of a *Gaussian*

[19]The computational advantage of kernel methods comes, in part, from relying on algorithms that only use inner products between inputs rather than the actual inputs. Shawe-Taylor and Cristianini (2004) also identify the modularity or re-usability of the learning algorithms as a computational advantage. Despite this, computational cost can still be an important constraint to consider. This topic will be discussed further in this thesis.

[20]This follows from the link between kernels and positive functions discussed before.

measure turns out to be the natural approach for studying the class of functions in the RKHS that best represent the phenomena studied in this work.

1.5 About this Thesis

The research presented in this work has the purpose of developing new machine learning tools, within the Gaussian Processes framework, for modelling spatiotemporal phenomena. We are particularly interested in methods that help studying malaria infections across population. This thesis is, in part, the result of a joint project with Makerere University, in Uganda, with the goal of modelling malaria spread and its relation with different environmental variables in that country. The needs and challenges of such an enterprise involve integrating different sources of information; being able to handle large scale data sets; and modelling non-Gaussian phenomena.

The structure of the work presented here is as follows. In Chapter 2, we present a review of the Gaussian process framework, with an emphasis on vector-valued regression, sparse approximations and approximate inference. The study of these topics is motivated by the needs and challenges mentioned above. In Chapter 3, we propose a sparse variant of the expectation propagation algorithm, which allows extending the sparse variational framework to non-Gaussian data. Departing from the results of Chapter 3, in Chapter 4, we explore a way to assimilate different data types and use dimensionality reduction to analyze data. We devote Chapter 5 to the study of malaria infections in Uganda. We apply the theoretical framework discussed in the previous chapters to the records of the Health Management Information System of Uganda. In Chapter 6, a monitoring system of malaria infection is proposed. Finally, in Chapter 7, we conclude with some final remarks and considerations for future work.

Chapters 3 and 4 are based on Andrade-Pacheco et al. (2014). Chapters 5 and 6 are based on Andrade-Pacheco et al. (2014) and Mubangizi et al. (2014). Chapter 7 is based on Andrade-Pacheco et al. (2015).

Chapter 2

Gaussian Processes

Gaussian processes provide a robust framework for probabilistic modelling. Their simplicity for doing inference has made of them one of the dominant methods for regression within the field of machine learning. They are commonly applied to *impulse-response* problems, but can also be extended to problems where the interrelation between different variables describes a closed loop. This feature is, indeed, desired for spatiotemporal modelling of multiple variables.

This flexibility does not come without a price. Computing scaling with large data sets has proved to be a strong limitation for GP models (Snelson and Ghahramani, 2006a). This problem is exacerbated when using *intrinsic correlations* for multiple variables. Different sparse approximation methods have been proposed to overcome this constraint, speed up the learning process and reduce the memory storage demands.

Non-Gaussian observation models can still be combined with a Gaussian *latent variable* and therefore managed under the GP framework (Neal, 1998). Whilst many phenomena can be satisfyingly modelled by assuming a Gaussian likelihood, this assumption cannot be sustained for many patterns of our surrounding reality. For example, continuity and symmetry assumptions are not always easy to justify when studying point patterns. In such cases, the multidimensional integration needed to compute the posterior distribution is intractable, and approximations are needed (Bishop, 2006). State of the art GP models with non-Gaussian likelihoods rely on a Markov chain Monte Carlo (MCMC) implementation (Adams et al., 2009a; Betancourt and Girolami, 2013; Knorr-Held and Rue, 2002; Korattikara et al., 2013), which can be strongly demanding in computing resources as well as time consuming. As an alternative to MCMC methods, the GP framework is equipped with approximate inference techniques which can be faster than MCMC and do not compromise the model's performance. *Variational Bayes* approximation (Hinton and van Camp, 1993;

Palmer et al., 2005) defines a lower bound to the model evidence or marginal likelihood p_y, by finding an approximation q_y that minimizes the Kullback-Leibler divergence between both, i.e., $\mathrm{KL}(q_y \| p_y)$ (see Appendix A.1). *Expectation Propagation* (Minka, 2001; Seeger, 2005) follows a similar approach to variational Bayes, but it is not defined as a lower bound, as it rather defines partial approximations using $\mathrm{KL}(p_y \| q_y)$ (the other way around from variational Bayes). *Laplace Approximation* (Williams and Barber, 1998) defines a Gaussian approximation based on the second order Taylor expansion around the posterior's mode (see Appendix A.2). In a more recent development, the *integrated nested Laplace approximation* computes a set of marginal posterior approximations by making use of the Laplace approximation in a stepwise approach (Rue et al., 2009).

In this chapter we will provide a formal introduction to Gaussian processes, followed by how they can be extended for multivariate cases using multiple output kernels. We will review a few sparse approximation methods. At last, a brief exposition of the *expectation propagation* algorithm will be presented. We first present a formal definition of some concepts that will be constantly used along our exposition.

2.1 Definitions

Definition 1 *(Hilbert space) A complete inner product space is called a Hilbert space.*

Definition 2 *(Scalar kernel) A function $K : \mathbb{S} \times \mathbb{S} \to \mathbb{C}$, where $\mathbb{S}$ is a non-empty abstract set, is a reproducing kernel of the Hilbert space $\mathcal{H}$ if and only if*

i) $\forall t \in \mathbb{S}$, $K(\cdot, t) \in \mathcal{H}$;

ii) $\forall t \in \mathbb{S}$ *and* $\forall \varphi \in \mathcal{H}$, $< \varphi, K(\cdot, t) >= \varphi(t)$.

Since we will be only working with processes in the real domain, from now on we will assume that $K : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$, with $\mathbb{S} \subseteq \mathbb{R}^q$ for some $q \in \mathbb{N}$.

Definition 3 *(Symmetric kernel) A kernel $K : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ is said to be symmetric if $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$, for $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{S}$.*

Definition 4 *(Positive semidefinite kernel) A kernel $K : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ is said to be positive semidefinite if the Gram matrix $\mathbf{K} = \left[K(\mathbf{x}_i, \mathbf{x}_j)\right]$ is positive semidefinite.*

Definition 5 *(Covariance kernel) We say that $K : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ is a covariance kernel if it is continuous, symmetric and positive semidefinite.*

Definition 6 *(Gaussian process) A stochastic process* $(f_s)_{s\in\mathbb{S}}$ *is said to be Gaussian if any finite linear combination of the real variables* $\mathbf{f} = (f_{s_1}, \ldots, f_{s_n})^\top$*,* $s_i \in \mathbb{S}$*, is a real Gaussian random variable.*

We will denote a Gaussian process as $(f_s) \sim \mathcal{GP}(\mathcal{M}, K)$, where $\mathcal{M}$ is a mean function and K is a covariance kernel function.

2.2 Gaussian Process Regression for Real-Valued Functions

Suppose we have a set of observations $\{(\mathbf{x}_i, y_i) |\, \mathbf{x}_i \in \mathbb{R}^q, y_i \in \mathbb{R}, i = 1, \ldots, n\}$, where $\mathbf{x}_i$ is regarded as the inputs of the output y_i, and we are interested in learning the functional relation between both. Hereafter, n will represent the number of observations and superindex q the dimensionality of each input. When convenient, a matrix notation $\mathbf{X} \in \mathbb{R}^{n\times q}$ and $\mathbf{y} \in \mathbb{R}^n$ will be used to represent the data instances.

In standard the regression case within the GP framework, the probability of an unknown function $f : \mathbb{R}^q \to \mathbb{R}$ is estimated from an observed data set. Since f is not observed, but is just an abstraction of the relation between inputs and outputs, its hypothetical realizations $\mathbf{f} = (f_{\mathbf{x}_1}, \ldots, f_{\mathbf{x}_n})^\top$ are regarded as *latent variables*. Usually, the elements of the set $\{y_i\}$ are assumed to be noisy realizations of f, such that

$$y_i = f_{\mathbf{x}_i} + \epsilon_i, \tag{2.1}$$

where $(f_{\mathbf{x}_i}) \sim \mathcal{GP}(\mathcal{M}, K)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then, applying *Bayes theorem*, the posterior distribution of $\mathbf{f}$ is computed as

$$p(\mathbf{f}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{f}|\mathbf{X})p(\mathbf{y}|\mathbf{f}, \mathbf{X})}{p(\mathbf{y}|\mathbf{X})}. \tag{2.2}$$

From the Gaussian assumption of the terms in the r.h.s. of equation (2.1), it follows that the likelihood term $p(\mathbf{y}|\mathbf{f}, \mathbf{X})$ and thus the posterior $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$ are also Gaussian. For such a model, $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$ has a simple analytical formulation, whose first and second moments are functions of the mean and covariance of the prior and the likelihood. Moreover, output predictions at a new input position $\mathbf{x}_*$ are computed consistently with the training data, through the predictive density $p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X})$. The mean and

variance of the predictive distribution are computed as

$$\langle y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}\rangle = \mathcal{M}(\mathbf{x}_*) + \mathbf{k_{*f}}\Big(\mathbf{K_{ff}} + \sigma^2\mathbf{I}\Big)^{-1}(\mathbf{y} - \mathcal{M}(\mathbf{X})), \tag{2.3}$$

$$\text{var}\,(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}) = k_{**} - \mathbf{k_{*f}}\Big(\mathbf{K_{ff}} + \sigma^2\mathbf{I}\Big)^{-1}\mathbf{k_{f*}} + \sigma^2, \tag{2.4}$$

where $k_{**} = K(\mathbf{x}_*, \mathbf{x}_*)$, $\mathbf{k_{f*}} = \mathbf{k_{*f}^\top} = K(\mathbf{X}, \mathbf{x}_*)$ and $\mathbf{K_{ff}} = K(\mathbf{X}, \mathbf{X})$. The font styles used to represent the evaluations of the kernel covariance function have the intention to emphasize the difference between the dimensions in each case ($k_{**} \in \mathbb{R}$, $\mathbf{k_{f*}} \in \mathbb{R}^{n\times 1}$ and $\mathbf{K_{ff}} \in \mathbb{R}^{n\times n}$). The shape of the mean function can also be parametrized and learnt from the data (Blight and Ott, 1975; O'Hagan and Kingman, 1978). However, in this work a simpler approach will be taken and it will be assumed $\mathcal{M}(\mathbf{x}) = 0 \;\; \forall \mathbf{x}$. This assumption does not necessarily have a negative impact in the performance of the model. The posterior mean is still defined using all the information provided in the training phase[1].

As mentioned before, an essential part of GP models is the covariance function. The kernel family can be sometimes defined *a priori*, depending on the characteristics of the process being modelled, but the parameters of the kernel (hyperparameters) should be learnt from the data as long as possible. This is usually done via hierarchical inference (Gelman et al., 2013). The approach followed here is to select the hyperparameters by maximizing the marginal likelihood $p(\mathbf{y}|\mathbf{X})$.

2.3 Gaussian Process Regression for Vector-Valued Functions

The use of real-valued random functions, although simple and flexible, can be inadequate or restrictive in some applications. For example, if the object of study is a motorcycle moving along a racing circuit, we might be interested in modelling variations in 2 (or 3) coordinate axis, as well as the lean angle. Of course, we could use an independent real-valued GP for each variable we track, but any relation between them would be neglected, and thus our correlation model would be flawed. Another example is studying the patterns of the kick drum and bass guitar in a rock song. Some studio-producing techniques exploit the relation between the frequencies of these instruments in order

[1] In some applications it might be adequate to use a mean function different from zero. For example, when modelling supernova light curves, Kim et al. (2013) avoid using a zero mean to prevent zero flux expectations in temporal regions with no data.

to enhance the sound[2]. The later example is case of *multiple task learning*, where the pattern of each instrument is a different task to learn. The first example can be thought as combination of *multiple output learning* and multiple task learning. The movement across each coordinate axis can be easily thought to be monitored by the same device (e.g., GPS), and so we have a function that is producing multiple outputs. However, the lean angle could be monitored by a different device and it could even be sampled at different time-points than the location[3].

Both kind of problems can be represented as a closed loop system, where none of the variables tracked is the input of one another, but they present some (symmetric) relation. In time series literature, this kind of problems are commonly treated in the family of VAR models (Quenouille, 1957), while in geostatistics literature *co-Kriging* generalizations are used (Matheron, 1982; Myers, 1982). The approach is similar in both cases, generalize the concepts of stationary random functions to the vector-valued case (Yaglom, 1986a,b).

Let h be a function that takes values in some d-dimensional Euclidean space $\mathcal{Y}$. The realizations of h can be thought as composed of the realization of d real-valued functions, each one related to a different output (task); i.e., $h_{\mathbf{x}} = (f_{\mathbf{x}}^1, \dots, f_{\mathbf{x}}^d)^\top$ for $f_{\mathbf{x}}^i : \mathbb{R}^q \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^q$. The corresponding mean vector and covariance matrix are given by

$$\langle h_{\mathbf{x}} \rangle = \Big(\langle f_{\mathbf{x}}^1 \rangle, \dots, \langle f_{\mathbf{x}}^d \rangle \Big)^\top, \tag{2.5}$$

$$\Big[\operatorname{cov}(h_{\mathbf{x}}, h_{\mathbf{z}})_{ij} \Big] = \Big[\operatorname{cov}(f_{\mathbf{x}}^i, f_{\mathbf{z}}^j) \Big]. \tag{2.6}$$

The diagonal elements of the correlation matrix $\Big[\operatorname{cov}(h_{\mathbf{x}}, h_{\mathbf{z}})_{ii} \Big]$ are just the covariance functions of the real-valued components. More interesting are the non-diagonal elements, which represent the *cross-covariance functions* between components. For multiple task learning problems the cross-covariance functions allow learning one process, which might be difficult or expensive to track, from another one, whose samples are more abundant or cheaper.

A formal generalization of learning theory with RKHS of vector-valued functions has been studied by Micchelli and Pontil (2004, 2005) and Baldassarre et al. (2012). It turns out that if $\mathcal{Y} \subseteq \mathbb{R}^d$, then the space of bounded linear operators $\mathcal{B}(\mathcal{Y})$, where

[2]Among the common tricks, they lower the bass when the kick drum attacks or add a pitched decay to the drum impulse or add a submix of both instruments to the overall song to ensure their sounds are tightened.

[3]The terms multiple output and multiple task learning belong to the *statistical learning theory*, but are equivalent to the *isotopic* and *heterotopic* concepts defined by Matheron (1982), when he introduced the *coregionalization model* for geostatistics.

the corresponding kernels take values, is the space of $d \times d$ matrices. This is analogue, and in fact equivalent, to the covariance function as defined in Equation (2.6). Álvarez et al. (2012) present a review of *separable kernels* that can be expressed as a sums of products between kernels for the input space $\mathbb{R}^q$ and kernels for the index set of output (task) components $\mathcal{J} = \{1, \dots, d\}$. We will follow this kernel construction. Let $K : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ and $B : \mathcal{J} \times \mathcal{J} \to \mathbb{R}$ be kernels for the input space and for the index set, respectively, then a separable multi-output kernel $\Gamma : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^{d \times d}$ can be formulated as

$$\begin{aligned} \Gamma(\mathbf{x}, \mathbf{z}) &= \Big[K(\mathbf{x}, \mathbf{z}) \times B(i, j)\Big] \\ &= K(\mathbf{x}, \mathbf{z}) \times \mathbf{B}, \end{aligned} \tag{2.7}$$

where

$$\mathbf{B} = \Big[B(i, j)\Big], \tag{2.8}$$

for $i, j \in \mathcal{J}$. In this formulation, kernel K has the same interpretation as any kernel on the input space of real-valued functions. In contrast, kernel B (and therefore the matrix $\mathbf{B}$) is interpreted as an encoder of the interactions among outputs (tasks). In geostatistics, a multivariate spatial dependence assumed to have an structure as in equation (2.7) is said to be *intrinsically coregionalized* (Helterbrand and Cressie, 1994). In such context, $\mathbf{B}$ is also known as *coregionalization matrix*.

Different encoders can be combined with kernels on the input space to construct flexible models. This gives rise to the *linear coregionalization models* (LCM), which use kernels defined as a linear combination of intrinsic kernels such that

$$\Gamma(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{Q} K_i(\mathbf{x}, \mathbf{z}) \times \mathbf{B}_i, \tag{2.9}$$

for some $Q \in \mathbb{N}$.

Hein and Bousquet (2004) showed that any kernel $\Gamma : \mathbb{S} \times \mathbb{S} \to \mathbb{R}^{d \times d}$ can be seen as a scalar kernel $\Gamma' : (\mathbb{S}, \mathcal{J}) \times (\mathbb{S}, \mathcal{J}) \to \mathbb{R}$. Therefore, if we increase the input dimensionality to include the task index, an equivalent model to (2.9) is given by

$$\Gamma'\Big((\mathbf{x}^\top, \jmath(\mathbf{x}))^\top, (\mathbf{z}^\top, \jmath(\mathbf{z}))^\top\Big) = \sum_{i=1}^{Q} K_i(\mathbf{x}, \mathbf{z}) \times B_i(\jmath(\mathbf{x}), \jmath(\mathbf{z})), \tag{2.10}$$

for $\jmath(\mathbf{x}), \jmath(\mathbf{z}) \in \mathcal{J}$.

Two final considerations are worth mentioning before moving to the next topic. First, an intrinsic covariance kernel (equation (2.7)) such that $\mathbf{B} = \mathbf{I}$ can be thought

of equivalent to fitting d independent real-valued GPs. Indeed, each GP would be independent, however the learning process would not. If we fitted d independent regression models, there would be d sets of hyperparameters to be learnt. In a model like (2.7) there is only one set of hyperparameters shared across all the data sets. A second consideration is related to the size of the covariance matrices used by these models. If each tasks has n_i training points, then the gram matrix of the kernel is of size $(\sum_{i=1}^{d} n_i) \times (\sum_{i=1}^{d} n_i)$. Roughly speaking, the size of the covariance matrix increases quadratically in the number of outputs. This leads us directly into the next section, where we will talk about how the size of the covariance matrix can result prohibiting for the application of Gaussian processes, and what alternatives can be implemented.

2.4 Sparse Approximations for Gaussian Process Regression

The computation of the predictive mean and predictive variance, in Equations (2.3) and (2.4), requires computing the matrix $(\mathbf{K_{ff}} + \sigma^2\mathbf{I})^{-1}$. Inverting a matrix is an operation usually recommended to avoid unless it is absolutely necessary (Higham, 2002). Unfortunately, this is one of those cases when it is needed to compute an inverse. For this framework to be feasible for large data sets, the burden of storing large matrices ($\mathcal{O}(n^2)$) and inverting them ($\mathcal{O}(n^3)$) has to be reduced in some way.

Sparse Gaussian processes are low rank approximations based on a small set of *inducing latent variables* $\mathbf{u} = (u_{\mathbf{z}_1}, \ldots, u_{\mathbf{z}_m})^\top$, associated to a set of *inducing inputs* $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)^\top$, where $\mathbf{Z}$ and $\mathbf{X}$ belong to the same domain and $m < n$ (Lawrence, 2007; Quiñonero Candela and Rasmussen, 2005; Seeger et al., 2003; Snelson and Ghahramani, 2006b; Titsias, 2009). Rather than computing the covariance between any pair of variables $f_{\mathbf{x}_i}$ and $f_{\mathbf{x}_j}$, sparse approximations induce their relation through their dependence on the elements of $\mathbf{u}$. Thus the computation of an $n \times n$ covariance matrix is no longer needed. The storage demands and computational complexity are reduced even when computing a predictive distribution. The relation between the latent variables $\mathbf{f}$ and $\mathbf{f}_*$ (associated to a new set of inputs $\mathbf{X}_*$) is also a generalization of their dependence on $\mathbf{u}$. In this section we will present some of the popular sparse models in the literature. The exposition is based mainly in the work of Quiñonero Candela and Rasmussen (2005) and Titsias (2009). For simplicity, we will assume real-valued functions only, rather than the vector valued case.

2.4.1 Exact Conditionals

Suppose a set of inducing inputs $\mathbf{Z}$ and the corresponding latent variables $\mathbf{u}$ is given. The exact expression for the conditionals of $\mathbf{f}$ and $\mathbf{f}_*$ on the inducing variables and all the input locations are the following

$$p(\mathbf{f}|\mathbf{u},\mathcal{X}) = \mathcal{N}(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \mathbf{K_{ff}} - \mathbf{Q_{ff}}), \tag{2.11}$$

$$p(\mathbf{f}_*|\mathbf{u},\mathcal{X}) = \mathcal{N}(\mathbf{K_{*u}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \mathbf{K_{**}} - \mathbf{Q_{**}}); \tag{2.12}$$

where $\mathcal{X} \subseteq \{\mathbf{Z}, \mathbf{X}, \mathbf{X}_*\}$ is the corresponding subset of inputs to each conditional, $\mathbf{K_{fu}} = \mathbf{K_{uf}^\top} = K(\mathbf{X},\mathbf{Z})$, $\mathbf{K_{uu}} = K(\mathbf{Z},\mathbf{Z})$, $\mathbf{K_{u*}} = \mathbf{K_{*u}^\top} = K(\mathbf{Z},\mathbf{X}_*)$, $\mathbf{K_{**}} = K(\mathbf{X}_*,\mathbf{X}_*)$, $\mathbf{Q_{ff}} = \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}$, and $\mathbf{Q_{**}} = \mathbf{K_{*u}}\mathbf{K_{uu}^{-1}}\mathbf{K_{u*}}$. The introduction of the inducing latent variables provides no benefit, if the exact model is used. Once the inducing latent variables are marginalized out for making output predictions, we will have exact predictive distribution with mean and variance given by Equations (2.3) and (2.4). However, as it will be exposed next, if the dependence on $\mathbf{u}$ does not use the full covariance structure, but an approximation, a reduction in the storage demand and complexity can be achieved.

2.4.2 Deterministic Training Conditional Approximation (DTC)

This model assumes a *deterministic conditional* approximation for the training set, but uses the exact conditional for the test set. The used conditionals are given by

$$q(\mathbf{f}|\mathbf{u},\mathcal{X}) = \mathcal{N}(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \mathbf{0}), \tag{2.13}$$

$$q(\mathbf{f}_*|\mathbf{u},\mathcal{X}) = p(\mathbf{f}_*|\mathbf{u},\mathcal{X}). \tag{2.14}$$

When the conditionals above are assumed, the joint prior distribution of $(\mathbf{f},\mathbf{f}_*)^\top$ implied by the model is

$$q(\mathbf{f},\mathbf{f}_*|\mathcal{X}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q_{ff}} & \mathbf{Q_{*f}} \\ \mathbf{Q_{f*}} & \mathbf{K_{**}} \end{bmatrix}\right). \tag{2.15}$$

The posterior distribution of $\mathbf{f}_*$, once the latent variables $\mathbf{u}$ have been marginalized out, is

$$q(\mathbf{f}_*|\mathbf{y},\mathcal{X}) = \mathcal{N}\left(\mathbf{Q_{*f}}\left(\mathbf{Q_{ff}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}, \mathbf{K_{**}} - \mathbf{Q_{*f}}\left(\mathbf{Q_{ff}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{Q_{f*}}\right). \tag{2.16}$$

All data instances are still used in this model. Yet this approximation achieves a computational complexity of $\mathcal{O}(nm^2)$, by using the low rank matrix $\mathbf{Q_{ff}}$ instead of $\mathbf{K_{ff}}$. Notice, however, that the entries of the covariance matrix depend on the type of instance and not only on the distance between inputs[4]. This means that the usual rules of marginalization of a collection of Gaussian variables are not followed. Thus, the initial definition of a Gaussian process (Definition 6) is not satisfied.

2.4.3 Fully Independent Training Conditional Approximation (FITC)

In this approximation, all test points are considered conditionally independent from each other. However the assumed variance of each one is the variance of the exact model. The conditional distribution for the test set is

$$q(\mathbf{f}|\mathbf{u}, \mathcal{X}) = \mathcal{N}(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \mathbf{D_{ff}}), \tag{2.17}$$

where $\mathbf{D_{ff}} = \text{diag}\,(\mathbf{K_{ff}} - \mathbf{Q_{ff}})$.

The conditional distribution for the test data can be defined in two ways. Either case affects the nature of the approximation, but not the complexity, which is also $\mathcal{O}(nm^2)$.

1. Exact test conditional:

$$q(\mathbf{f}_*|\mathbf{u}, \mathcal{X}) = p(\mathbf{f}_*|\mathbf{u}, \mathcal{X}), \tag{2.18}$$

   which implies the following joint prior:

$$q(\mathbf{f}, \mathbf{f}_*|\mathcal{X}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q_{ff}} + \mathbf{D_{ff}} & \mathbf{Q_{*f}} \\ \mathbf{Q_{f*}} & \mathbf{K_{**}} \end{bmatrix}\right). \tag{2.19}$$

2. Independent test conditional:

$$q(\mathbf{f}_*|\mathbf{u}, \mathcal{X}) = \mathcal{N}(\mathbf{K_{*u}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \mathbf{D_{**}}), \tag{2.20}$$

[4] The prior covariances used depend on whether the data instances are considered as part of the training or test set.

which implies the following prior:

$$q(\mathbf{f}, \mathbf{f}_*|\mathcal{X}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q_{ff}} + \mathbf{D_{ff}} & \mathbf{Q_{*f}} \\ \mathbf{Q_{f*}} & \mathbf{Q_{**}} + \mathbf{D_{**}} \end{bmatrix}\right). \tag{2.21}$$

In the first case, the covariance entries shown in Equation (2.19)) depend on the type of instance (training or test). Hence, the model does not have a consistent joint Gaussian distribution, just like in DTC approximation. Nevertheless, if the test data is a single point, the exact test conditional and the independent test conditional are equivalent (see Equation (2.21)). In both cases the model has a consistent joint Gaussian distribution. The posterior distribution for a single point f_* is expressed as

$$q(f_*|\mathbf{y}, \mathcal{X}) = \mathcal{N}\left(\mathbf{Q_{*f}}\left(\mathbf{Q_{ff}} + \mathbf{D_{ff}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}, \mathbf{K_{**}} - \mathbf{Q_{*f}}\left(\mathbf{Q_{ff}} + \mathbf{D_{ff}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{Q_{f*}}\right). \tag{2.22}$$

2.4.4 Selection of the Inducing Inputs

The performance of a sparse model is strongly dependent on the selection of the inducing inputs $\mathbf{Z}$. If they are constrained to be part of the training set (i.e., $\mathbf{z}_i \in \mathbf{X}$), choosing them becomes a combinatorial optimization problem. Nevertheless, their selection can be turned into a simpler continuous optimization problem if the elements $\mathbf{z}_i$ are allowed to be any point in $\mathbb{R}^q$. In this case, the inducing set can be found by maximizing the marginal likelihood with respect to $\mathbf{Z}$. An advantage of proceeding in this way is that the model hyperparameters can be learnt at the same time as the inducing inputs. For the models described above, the marginal log-likelihood can be formulated as

$$\begin{aligned} \log q(\mathbf{y}|\mathcal{X}) &= \log \iint p(\mathbf{y}|\mathbf{f}, \mathcal{X}) q(\mathbf{f}|\mathbf{u}, \mathcal{X}) p(\mathbf{u}|\mathcal{X}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f} \\ &= \log \int p(\mathbf{y}|\mathbf{f}, \mathcal{X}) q(\mathbf{f}|\mathcal{X}) \mathrm{d}\mathbf{f} \\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{Q_{f,f}} + \mathbf{\Lambda}| - \frac{1}{2}\mathbf{y}^\top(\mathbf{Q_{f,f}} + \mathbf{\Lambda})^{-1}\mathbf{y}, \end{aligned} \tag{2.23}$$

where the shape of $\mathbf{\Lambda}$ depends on the type of approximation, so that

$$\mathbf{\Lambda}_{DTC} = \sigma^2\mathbf{I}, \tag{2.24}$$

$$\mathbf{\Lambda}_{FITC} = \mathrm{diag}\left(\mathbf{K_{ff}} - \mathbf{Q_{ff}}\right) + \sigma^2\mathbf{I}. \tag{2.25}$$

2.4.5 Probabilistic Variational Sparse GP Approximation

Titsias (2009) introduced a variational method that jointly selects the inducing inputs and the hyperparameters by maximizing a lower bound to the exact marginal likelihood. A rigorous lower bound on the marginal log-likelihood allows joint optimization of the inducing inputs and hyperparameters without overfitting. This model approximates the true predictive distribution given by

$$\begin{aligned} p(\mathbf{y}_*|\mathbf{y},\mathcal{X}) &= \int p(\mathbf{y}_*|\mathbf{f},\mathcal{X})p(\mathbf{f}|\mathbf{y},\mathcal{X})\mathrm{d}\mathbf{f} \\ &= \iint p(\mathbf{y}_*|\mathbf{u},\mathbf{f},\mathcal{X})p(\mathbf{f}|\mathbf{u},\mathbf{y},\mathcal{X})p(\mathbf{u}|\mathbf{y},\mathcal{X})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f}, \end{aligned} \tag{2.26}$$

with an approximation defined as

$$q(\mathbf{y}_*|\mathcal{X}) = \int p(\mathbf{y}_*|\mathbf{u},\mathcal{X})\phi(\mathbf{u})\mathrm{d}\mathbf{u} \triangleq \int q(\mathbf{y}_*,\mathbf{u}|\mathcal{X})\mathrm{d}\mathbf{u}, \tag{2.27}$$

where $\phi(\mathbf{u})$ is a *free variational Gaussian distribution* $\mathcal{N}(\mathbf{u}|\mathbf{m},\mathbf{C})$.

The selection of the parameters $\mathbf{m}$ and $\mathbf{C}$, as well as the inducing set $\mathbf{u}$, is done by minimizing $\mathrm{KL}\left(p(\mathbf{f}|\mathbf{u},\mathcal{X})\phi(\mathbf{u})\|p(\mathbf{f},\mathbf{u}|\mathbf{y},\mathcal{X})\right)$, which, following Appendix A.1, is equivalent to maximizing the lower bound

$$\mathcal{L}_T(\mathbf{u},\phi) = \int p(\mathbf{f}|\mathbf{u},\mathcal{X})\phi(\mathbf{u})\log\frac{p(\mathbf{y}|\mathbf{f},\mathcal{X})p(\mathbf{u}|\mathcal{X})}{\phi(\mathbf{u})}\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u}. \tag{2.28}$$

The distribution $\hat{\phi}$ that maximizes $\mathcal{L}_T(\mathbf{u},\phi)$ has parameters $\hat{\mathbf{m}}$ and $\hat{\mathbf{C}}$, expressed as

$$\hat{\mathbf{m}} = \sigma^{-2}\mathbf{K}_{\mathbf{uu}}^{-1}\hat{\mathbf{\Sigma}}^{-1}\mathbf{K}_{\mathbf{uf}}\mathbf{y}, \tag{2.29}$$

$$\hat{\mathbf{C}} = \mathbf{K}_{\mathbf{uu}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{K}_{\mathbf{uu}}; \tag{2.30}$$

where

$$\hat{\mathbf{\Sigma}} = \mathbf{K}_{\mathbf{uu}} + \sigma^{-2}\mathbf{K}_{\mathbf{uf}}\mathbf{K}_{\mathbf{fu}}. \tag{2.31}$$

After optimization w.r.t. ϕ, the expression of the lower bound is the following

$$\mathcal{L}_T = \log\mathcal{N}(\mathbf{y}|\mathbf{0},\sigma^2\mathbf{I}+\mathbf{Q}_{\mathbf{ff}}) - \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{K}_{\mathbf{ff}}-\mathbf{Q}_{\mathbf{ff}}\right). \tag{2.32}$$

The shape of the bound $\mathcal{L}_T$ has exactly the shape of the marginal likelihood in DTC, but with an additional trace term. This additional term is interpreted as a

correction term on the DTC approximation that penalizes the likelihood depending on how different the approximation from the true variance is.

The predictive distribution of a new observation y_* is expressed as

$$q(y_*|\mathbf{y}, \mathcal{X}) = \mathcal{N}(y_*|\mathbf{k}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\hat{\mathbf{m}}, k_{**} - \mathbf{k}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}*} + \mathbf{k}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\hat{\mathbf{C}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}*}). \tag{2.33}$$

So far we have talked about learning multiple processes within the GP framework and the complexities involved with large covariance matrices. Another important point to consider in this review is that spatiotemporal processes in nature cannot always be thought of as Gaussian. The next section is motivated by the need to implement models with non-Gaussian noise. We will revisit the expectation propagation algorithm for doing approximate inference.

2.5 Approximate Inference with EP

Non-Gaussian likelihoods can also be modelled within the GP framework, if they are assumed to be a convenient function over the latent variables $g(f_{\mathbf{x}_i})$. For example, in binary classification, where we take $y_i \in \{0, 1\}$, the realizations of a Gaussian process are normally mapped through a *squashing function* $g : \mathbb{R} \mapsto (0, 1)$ to provide a set of probabilities $\{\pi_i = g(f_{\mathbf{x}_i})|i = 1, \ldots, n\}$, which can then be used as parameters of a Bernoulli likelihood $p(y_i|f_{\mathbf{x}_i}) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$. This is an analogue approach to the one followed by *generalized linear models* in a parametric setting (see appendix B).

Such a non-linear transformation over $f_{\mathbf{x}_i}$ renders exact inference in the resulting model intractable. This led Barber and Williams (1997) to consider the Laplace approximation (Appendix A.2) and Gibbs and MacKay (2000) to adopt a variational lower bound from Jaakkola and Jordan (1996)[5] to make progress. The more standard variational approach (often known as variational inference), based on minimizing the Kullback-Leibler divergence between an approximation and the true posterior density (Appendix A.1), has also been proposed for non-Gaussian data. Seeger (2004) considered this approximation for classification and Tipping and Lawrence (2003) extended the relevance vector machine[6] to heavy tailed data. However, as shown empirically by Kuss and Rasmussen (2005), for the case of classification, standard application of variational inference to *sub-Gaussian* likelihoods can lead to very poor approximations of the

[5]This variational lower bound exploited the log-convexity of a sigmoidal squashing function, but does not follow the standard approach to variational inference.

[6]A sparse Bayesian regression model that can also be expressed as a GP with a degenerate covariance.

marginal likelihood. Instead, the expectation propagation algorithm (Minka, 2001; Opper and Winther, 2000) is generally preferred. Both EP and its variants have been applied to likelihoods that allow semi-supervised learning (Lawrence and Jordan, 2005), ordinal regression (Chu and Ghahramani, 2005) and binary classification (Kuss and Rasmussen, 2005). However, its application in the context of heavy tailed likelihoods is generally more involved (Jylänki et al., 2011).

2.5.1 Standard EP with Site Gaussian Approximations

For Gaussian process models, EP combines a Gaussian prior $p(\mathbf{f}|\mathbf{X})$ with a set of site approximations to the likelihood[7] $\{\ell_i(f_{\mathbf{x}_i}) \approx p(y_i|f_{\mathbf{x}_i})|i = 1, \ldots, n\}$. This results in an approximation to the posterior density of $\mathbf{f}$ given by

$$q(\mathbf{f}|\mathbf{y}, \mathbf{X}) = \frac{1}{Z_{EP}} p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{n} \ell_i(f_{\mathbf{x}_i}), \tag{2.34}$$

where Z_{EP} is the normalizing constant of $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$ (see Williams and Rasmussen (2006) for notation).

Let the factors ℓ_i in Equation (2.34) be un-normalized Gaussians up to a constant $\tilde{Z}_i$, with parameters $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$. We can write them as $\ell_i(f_{\mathbf{x}_i}) = \ell_i(f_{\mathbf{x}_i}|\tilde{\mu}_i, \tilde{\sigma}_i^2, \tilde{Z}_i) \triangleq \tilde{Z}_i \mathcal{N}(f_{\mathbf{x}_i}|\tilde{\mu}_i, \tilde{\sigma}_i^2)$. Overall, these factors are combined to provide a Gaussian-like approximation to the likelihood

$$p(\mathbf{y}|\mathbf{f}) \approx \tilde{\boldsymbol{Z}} \times \mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{2.35}$$

for some constant $\tilde{\boldsymbol{Z}}$.

To ease notation, while explaining EP algorithm, the site approximations will be reparametrized to use the natural parameters $\{\tilde{\tau}_i, \tilde{\nu}_i\}$, rather than the moment parameters $\{\tilde{\mu}_i, \tilde{\sigma}_i^2\}$. We will use $\mathcal{N}(f_{\mathbf{x}_i}|\tilde{\boldsymbol{\theta}}_i)$ instead of $\mathcal{N}(f_{\mathbf{x}_i}|\tilde{\mu}_i, \tilde{\sigma}_i^2)$, where $\tilde{\boldsymbol{\theta}}_i = (\tilde{\tau}_i, \tilde{\nu}_i)^\top$, $\tilde{\tau}_i = \tilde{\sigma}_i^{-2}$ and $\tilde{\nu}_i = \tilde{\sigma}_i^{-2}\tilde{\mu}_i$.

The parameters $\{\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i\}$ are adjusted through an iterative approach, usually starting with $\ell_i(f_{\mathbf{x}_i}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i) = 1 \; \forall \, i$, until convergence is achieved (Seeger, 2005). In each step, factor $\ell_i(f_{\mathbf{x}_i}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i)$ in $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$ is updated, while fixing $\ell_j(f_{\mathbf{x}_j}|\tilde{Z}_j, \tilde{\boldsymbol{\theta}}_j) \; \forall j \neq i$, in order to make the global approximation $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$ as close as possible to a distribution given by

$$\hat{q}(\mathbf{f}|\mathbf{y}, \mathbf{X}) \propto p(y_i|f_{\mathbf{x}_i}) p(\mathbf{f}|\mathbf{X}) \prod_{j \neq i} \ell_j(f_{\mathbf{x}_j}|\tilde{Z}_j, \tilde{\boldsymbol{\theta}}_j). \tag{2.36}$$

[7] EP can be defined in a more general way, but we will only use this definition for simplicity.

In other words, $\ell_i(f_{\mathbf{x}_i}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i)$ is forced to behave as close as possible to $p(y_i|f_{\mathbf{x}_i})$ in $\hat{q}(\mathbf{f}|\mathbf{y}, \mathbf{X})$. Thus, each step consists of solving

$$\underset{\ell_i(f_{\mathbf{x}_i}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i)}{\operatorname{argmin}} \; \mathrm{KL}\left(\hat{q}(\mathbf{f}|\mathbf{y}, \mathbf{X}) \| q(\mathbf{f}|\mathbf{y}, \mathbf{X})\right). \tag{2.37}$$

Since $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$ is Gaussian, the previous KL divergence is minimized when the first and second moments of both distributions match (Kuss and Rasmussen, 2005), i.e., when

$$\langle(\mathbf{f}, \mathbf{f}\mathbf{f}^\top)^\top\rangle_{\hat{q}} = \langle(\mathbf{f}, \mathbf{f}\mathbf{f}^\top)^\top\rangle_{q}. \tag{2.38}$$

Expectations in the equation above involve integrating over $f_{\mathbf{x}_j}$, for $j = 1, \ldots, n$. Since only the i-th likelihood factor in $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$ is being modified and it is independent from the rest, it is easier to marginalize and work just on $f_{\mathbf{x}_i}$. In fact, the marginals for $j \neq i$ are the same in $\hat{q}(\mathbf{f}|\mathbf{y}, \mathbf{X})$ and $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$.

Let the cavity distribution have the form

$$q_{-i}(f_{\mathbf{x}_i}) \propto \int p(\mathbf{f}|\mathbf{X}) \prod_{j \neq i} \ell_j(f_{\mathbf{x}_j}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_j) \mathrm{d}f_{\mathbf{x}_j}. \tag{2.39}$$

If $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$ has Gaussian marginals $\mathcal{N}(f_{\mathbf{x}_i}|\boldsymbol{\theta}_i)$, then $q_{-i}(f_{\mathbf{x}_i}) = \mathcal{N}(f_{\mathbf{x}_i}|\boldsymbol{\theta}_{-i})$, with $\boldsymbol{\theta}_{-i} = \boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i$. Therefore, we have that

$$\begin{aligned} \ell_i(f_{\mathbf{x}_i}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i) q_{-i}(f_{\mathbf{x}_i}) &= \tilde{Z}_i \mathcal{N}(f_{\mathbf{x}_i}|\tilde{\boldsymbol{\theta}}_i) \mathcal{N}(f_{\mathbf{x}_i}|\boldsymbol{\theta}_{-i}) \\ &\propto \tilde{Z}_i \mathcal{N}(f_{\mathbf{x}_i}|\tilde{\boldsymbol{\theta}}_i + \boldsymbol{\theta}_{-i}). \end{aligned} \tag{2.40}$$

For Equation (2.38) to be satisfied, we need the moments of the distribution shown en Equation (2.40) be equal to the moments of $p(y_i|f_{\mathbf{x}_i}) q_{-i}(f_{\mathbf{x}_i})$. Lets call these moments $\hat{\boldsymbol{\theta}}_i$. Then, it is needed that $\hat{\boldsymbol{\theta}}_i = \tilde{\boldsymbol{\theta}}_i + \boldsymbol{\theta}_{-i}$, and the new natural parameters ${}_u\tilde{\boldsymbol{\theta}}_i$ are calculated as

$${}_u\tilde{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{-i}. \tag{2.41}$$

In addition, to have a normalized global approximation and since $\hat{q}(\mathbf{f}|\mathbf{y}, \mathbf{X})$ is un-normalized, it is required that the zero-th moments match in both distributions. This means that the normalizing constant $\tilde{Z}_i$ is chosen so that

$$\int p(y_i|f_{\mathbf{x}_i}) q_{-i}(f_{\mathbf{x}_i}) \mathrm{d}f_{\mathbf{x}_i} = \int \ell_i(f_{\mathbf{x}_i}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i) q_{-i}(f_{\mathbf{x}_i}) \mathrm{d}f_{\mathbf{x}_i}. \tag{2.42}$$

Once convergence has been achieved, the marginal log-likelihood approximation is calculated according to

$$\begin{aligned}
\log p(\mathbf{y}) \approx \log Z_{EP} &= \log \int p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{n} \ell_i(f_{\mathbf{x}_i}|\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i)\mathrm{d}\mathbf{f} \\
&= \log \int \mathcal{N}(\mathbf{f}|\boldsymbol{\theta}_0) \prod_{i=1}^{n} \mathcal{N}(f_{\mathbf{x}_i}|\tilde{\boldsymbol{\theta}}_i)\tilde{Z}_i\mathrm{d}\mathbf{f} \\
&= \log \int \mathcal{N}(\mathbf{f}|\boldsymbol{\theta}_0)\mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\theta}}) \prod_{i=1}^{n} \tilde{Z}_i\mathrm{d}\mathbf{f} \\
&= \log \int \mathcal{N}(\mathbf{f}|\boldsymbol{\theta}_0 + \tilde{\boldsymbol{\theta}})Z_0^{-1} \prod_{i=1}^{n} \tilde{Z}_i\mathrm{d}\mathbf{f} \\
&= \log Z_0^{-1} + \sum_{i=1}^{n} \log \tilde{Z}_i,
\end{aligned} \tag{2.43}$$

where $\boldsymbol{\theta}_0$ are the parameters of the prior distribution and Z_0^{-1} is the normalizing constant of $\mathcal{N}(\mathbf{f}|\boldsymbol{\theta}_0 + \tilde{\boldsymbol{\theta}})$.

2.5.2 EP-FITC

In the standard EP algorithm, global approximation parameters are calculated as $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \tilde{\boldsymbol{\theta}}$. This simple expression (in natural parameters), implies the inversion of an $n \times n$ matrix for calculating the new variance (moment parameters). Naish-Guzman and Holden (2008) introduced an sparse approach for EP, based on the FITC approximation, which reduces complexity to $\mathcal{O}(nm^2)$. This approximation works under the same principles of the standard EP. Sparsity is achieved by substituting $p(\mathbf{f}|\mathbf{X})$ in equation (2.34) with its FITC approximation (see in equation (2.17)). Not only the posterior distribution is computed in a cheaper way with this approach, also the computing complexity when calculating the predictive distribution is reduced.

In the standard EP, the predictive distribution for a new observation y_* is estimated as

$$\begin{aligned}
p(y_*|\mathbf{y}, \mathcal{X}) &= \int p(y_*|f_*, \mathcal{X})p(f_*|\mathbf{y}, \mathcal{X})\mathrm{d}f_* \\
&= \int p(y_*|f_*, \mathcal{X}) \int p(f_*|\mathbf{f}, \mathcal{X})p(\mathbf{f}|\mathbf{y}, \mathcal{X})\mathrm{d}\mathbf{f}\mathrm{d}f_* \\
&\approx \int p(y_*|f_*, \mathcal{X}) \int p(f_*|\mathbf{f}, \mathcal{X})q(\mathbf{f}|\mathbf{y}, \mathcal{X})\mathrm{d}\mathbf{f}\mathrm{d}f_*,
\end{aligned} \tag{2.44}$$

where $q(\mathbf{f}|\mathbf{y}, \mathcal{X})$ is the EP approximation. Under the EP-FITC approach the predictive distribution is estimated as

$$\begin{aligned} p(y_*|\mathbf{y}, \mathcal{X}) &= \int p(y_*|f_*, \mathcal{X}) p(f_*|\mathbf{y}, \mathcal{X}) \mathrm{d}f_* \\ &\approx \int p(y_*|f_*, \mathcal{X}) \int p(f_*|\mathbf{u}, \mathcal{X}) q(\mathbf{u}|\mathbf{y}, \mathcal{X}) \mathrm{d}\mathbf{u} \mathrm{d}f_* \\ &\approx \int p(y_*|f_*, \mathcal{X}) \int p(f_*|\mathbf{u}, \mathcal{X}) \int q(\mathbf{u}|\mathbf{f}, \mathcal{X}) q(\mathbf{f}|\mathbf{y}, \mathcal{X}) \mathrm{d}\mathbf{f} \mathrm{d}\mathbf{u} \mathrm{d}f_*, \end{aligned} \tag{2.45}$$

where

$$\begin{aligned} &q(\mathbf{f}|\mathbf{y}, \mathcal{X}) \equiv \text{ sparse EP approximation to } p(\mathbf{f}|\mathbf{y}, \mathcal{X}), \\ &q(\mathbf{u}|\mathbf{f}, \mathcal{X}) \propto q(\mathbf{f}|\mathbf{u}, \mathcal{X}) p(\mathbf{u}|\mathcal{X}), \\ &q(\mathbf{f}|\mathbf{u}, \mathcal{X}) \equiv \text{ FITC training conditional, and} \\ &(u_{\mathbf{z}}) \sim \mathcal{GP}. \end{aligned} \tag{2.46}$$

2.6 Final Comments

In this chapter we have reviewed different types of models, each one focused on solving a particular problem: vector-valued models for handling multiple outputs; sparse approximations for dealing with large data sets; and approximate inference for estimating non-tractable posterior distributions. All these models have been discussed by separate. In the next chapter, we will work approaches for combining these methods. We will propose a sparse variant of the EP algorithm and compare it with other methods. In Chapter 4, we will move forward and explore possible extensions of this new algorithm in a dimensionality reduction context.

Chapter 3

Variational Inference and EP

The study of spatiotemporal processes requires modelling assumptions beyond the Gaussian noise. In some cases, the variable to measure can be discrete or be constrained to a specific range of values. In other cases, instead of measuring the level of a variable, there might be a need for estimating the probability of an event's occurrence across time and space. The last describes a *point process*, which turns out to be essential to the field of spatial statistics (Diggle, 2003; Møller and Waagepetersen, 2007). We will not provide a thorough discussion on Poisson processes, as it would deviate us from the main topic. Instead, in Appendix, C we provide an introduction to these processes and show alternatives to implement them with the aid of a GP.

In Chapter 2, we reviewed approximation methods for large data sets (Section 2.4) and approximation methods for non-Gaussian data (Section 2.5). It is desirable to integrate both types of approximation into a single one. In this chapter, we move towards this goal: a single approximation method able to handle large data sets and non-Gaussian likelihoods. The need for an approximation comes from the fact that, in general, Bayesian inference is analytically intractable when using transformations on a Gaussian process to handle models of the type of GLM (Appendix B), like Poisson processes (Adams et al., 2009b).

3.1 Variational Lower Bound Recap

As we saw in Section 2.4, Titsias (2009) introduced a variational approximation to the regression problem that resulted in the lower bound formulated in Equation (2.32), repeated here

$$\mathcal{L}_T = \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{Q_{ff}} + \sigma^2\mathbf{I}\right) - \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{K_{ff}} - \mathbf{Q_{ff}}\right). \tag{3.1}$$

The first term in the r.h.s. of Equation (3.1) corresponds to the DTC likelihood approximation, where all the uncertainty is derived from the inducing latent variables $\mathbf{u}$ and propagated through $K(\cdot, \mathbf{u})$. The second term is a regularizer that penalizes using $\mathbf{Q_{ff}}$ instead of $\mathbf{K_{ff}}$, depending on how much their diagonals differ from each other. Large values of $\mathrm{tr}\left(\mathbf{K_{ff}} - \mathbf{Q_{ff}}\right)$ have to be compensated with a large variance or with kernel hyperparameters that give smooth functions. This means that this trace term is preventing overfitting. Notice also that if $\mathbf{Q_{ff}}$ and $\mathbf{K_{ff}}$ were the same, the exact model would be recovered.

The variational lower bound $\mathcal{L}_T$ is analytically tractable as long as a Gaussian likelihood is used. A different assumption would require an approximation, but then the complexity of the learning algorithm could be bounded by the complexity of the approximation if it is higher than $\mathcal{O}(nm^2)$. Our interest now is to extend the sparse variational regression model to the case of non-Gaussian likelihoods. We are also interested in an EP-type approximation due to its empirically proved performance (Kuss and Rasmussen, 2005; Vanhatalo et al., 2010). Although the sparse variant EP-FITC (Section 2.5.2) has the same complexity of the variational approach, it is not compatible with our goal, for it is not based on a lower bound to the marginal likelihood. In the next section, we will derive a variant of the EP algorithm based on the DTC approximation. Later on, we will show how this algorithm fits into the variational framework.

3.2 EP-DTC

To combine the EP approximation with the variational lower bound in equation (3.1), we first propose a derivation of the EP algorithm based on the DTC approximation (see Equation (2.13)). We refer to this algorithm as EP-DTC. Let the dependence of $\mathbf{f}$ on the inducing inputs $\mathbf{u}$ be defined deterministically according to

$$\mathbf{f} = \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \tag{3.2}$$

where $p(\mathbf{u}|\mathbf{Z}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K_{uu}})$. From this assumption, it follows that the marginal distribution of $\mathbf{f}$, which will be used as our prior information, is given by

$$q(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{Q_{ff}}). \tag{3.3}$$

As in standard EP (Section 2.5.1), the site approximations to the likelihood factors $\{\ell_i(f_{\mathbf{x}_i}) \approx p(y_i|f_{\mathbf{x}_i})|i = 1, \ldots, n\}$ are un-normalized Gaussians with moment parameters

$\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$. Thus, the overall likelihood approximation is given by

$$p(\mathbf{y}|\mathbf{f}) \approx \tilde{\mathbf{Z}} \times \mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{3.4}$$

for some constant $\tilde{\mathbf{Z}}$.

The gain of the sparse approximation depends on formulating the computation of the posterior moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in an efficient way. We will make these computations depend on the inversion of a $m \times m$ covariance matrix, rather than on one of size $n \times n$, for $m < n$. It can be proved[1] that the combination of the prior distribution in Equation (3.3) with the likelihood in Equation (3.4) yields a posterior distribution of $\mathbf{f}$ with parameters

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right), \tag{3.5}$$

$$\boldsymbol{\Sigma} = \left(\mathbf{Q}_{\mathbf{ff}}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}. \tag{3.6}$$

Now, by applying the matrix inversion lemma, the posterior variance is computed with a computational complexity of $\mathcal{O}(m^2 n)$ as follows

$$\begin{aligned}
\boldsymbol{\Sigma} &= \left(\left(\mathbf{K}_{\mathbf{fu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}}\right)^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1} \\
&= \tilde{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}}\left(\mathbf{K}_{\mathbf{fu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}} + \tilde{\boldsymbol{\Sigma}}\right)^{-1}\tilde{\boldsymbol{\Sigma}} \\
&= \mathbf{K}_{\mathbf{fu}}\left(\mathbf{K}_{\mathbf{uu}} + \mathbf{K}_{\mathbf{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K}_{\mathbf{fu}}\right)^{-1}\mathbf{K}_{\mathbf{uf}} \\
&= \mathbf{K}_{\mathbf{fu}}(\mathbf{L}\mathbf{L}^{\top})^{-1}\mathbf{K}_{\mathbf{uf}},
\end{aligned} \tag{3.7}$$

where $\mathbf{L} \in \mathbb{R}^{m \times m}$ is the Cholesky decomposition of $\mathbf{K}_{\mathbf{uu}} + \mathbf{K}_{\mathbf{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K}_{\mathbf{fu}}$.

In our sparse formulation, the procedure for updating the parameters of the site approximations remains the same as in standard EP. What changes is the computation of the posterior parameters, which now depends on the factorization of the covariance matrix given in Equation (3.7). As in Section 2.5.1, to simplify the notation we will explain these updates based on the natural parameters $\{\tilde{\tau}_i, \tilde{\nu}_i\}$, rather than on the moment parameters $\{\tilde{\mu}_i, \tilde{\sigma}_i^2\}$. Suppose that, after updating the i-th site approximation,

[1] This is consequence of the fact that both the prior and the likelihood approximation have a Gaussian shape.

the natural parameters change by $\Delta\tilde{\tau}_i$ and $\Delta\tilde{\nu}_i$. Let

$$\mathbf{E} = \tilde{\mathbf{\Sigma}}^{-1} + \Delta\tilde{\tau}_i \mathbf{e}_i \mathbf{e}_i^\top, \tag{3.8}$$

$$\mathbf{E}^{-1} = \tilde{\mathbf{\Sigma}} - \frac{\tilde{\tau}_i^2 \Delta\tilde{\tau}_i}{1 + \tilde{\tau}_i \Delta\tilde{\tau}_i} \mathbf{e}_i \mathbf{e}_i^\top, \tag{3.9}$$

where $\mathbf{e}_i$ is the i-th canonical basis vector of $\mathbb{R}^n$. Then, the updates of the posterior variance can be computed as

$$\begin{aligned} {}_u\mathbf{\Sigma} &= \left(\left(\mathbf{K_{fu}} \mathbf{K_{uu}^{-1}} \mathbf{K_{uf}} \right)^{-1} + \mathbf{E} \right)^{-1} \\ &= \mathbf{K_{fu}} (\mathbf{K_{uu}} + \mathbf{K_{uf}} \mathbf{E} \mathbf{K_{fu}})^{-1} \mathbf{K_{uf}} \\ &= \mathbf{K_{fu}} \left(\mathbf{L}\mathbf{L}^\top + \mathbf{k}_i \Delta\tilde{\tau}_i \mathbf{k}_i^\top \mathbf{K_{fu}} \right)^{-1} \mathbf{K_{uf}} \\ &= \mathbf{K_{fu}} \left({}_u\mathbf{L} \;\; {}_u\mathbf{L}^\top \right)^{-1} \mathbf{K_{uf}}, \end{aligned} \tag{3.10}$$

where $\mathbf{k}_i$ is the i-th column of $\mathbf{K_{uf}}$ and ${}_u\mathbf{L}$ is the Cholesky decomposition of $\mathbf{L}\mathbf{L}^\top + \mathbf{k}_i \Delta\tilde{\tau}_i \mathbf{k}_i^\top \mathbf{K_{fu}}$. Finally, the update of $\boldsymbol{\mu}$ is be computed as

$$\begin{aligned} {}_u\boldsymbol{\mu} &= {}_u\mathbf{\Sigma} \left(\mathbf{\Sigma}^{-1} \boldsymbol{\mu} + \Delta\tilde{\nu}_i \mathbf{e}_i \right) \\ &= {}_u\mathbf{\Sigma} \left(\left({}_u\mathbf{\Sigma}^{-1} - \Delta\tilde{\tau}_i \mathbf{e}_i \mathbf{e}_i^\top \right) \boldsymbol{\mu} + \Delta\tilde{\nu}_i \right) \\ &= \boldsymbol{\mu} + {}_u\mathbf{\Sigma} \left(\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i \mu_i \right) \mathbf{e}_i \\ &= \boldsymbol{\mu} + \left(\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i \mu_i \right) \; {}_u\mathbf{s}_i, \end{aligned} \tag{3.11}$$

where ${}_u\mathbf{s}_i$ is the i-th column of ${}_u\mathbf{\Sigma}$.

The derivation of this algorithm is analogue to EP-FITC (Naish-Guzman and Holden, 2008). However, we expect EP-DTC to be less competitive than the former. The reasons being the same as why FITC performs better that DTC (Snelson and Ghahramani, 2006a). In addition, EP-DTC tends to be less stable due to the comparatively weaker diagonal of the covariance matrix[2]. Nevertheless, our initial interest was not just to derive a formulation equivalent to DTC, but to extend the variational framework with a non-Gaussian approximation. In the next section we address this task.

[2]Remember that FITC uses the diagonal values of the actual covariance matrix $\mathbf{K_{ff}}$, while in DTC the diagonal values of the covariance matrix are estimated using $\mathbf{K_{fu}} \mathbf{K_{uu}^{-1}} \mathbf{K_{uf}} + \sigma^2 \mathbf{I}$.

3.3 EP in a Lower Bound Approximation

Assume we already have an optimal EP-DTC approximation of the form of equation (3.4). Following Titsias (2009) in using Jensen's inequality to define a lower bound on the logarithm of $p(\mathbf{y}|\mathbf{X})$, we see that

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &= \log \iint p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u},\mathcal{X})p(\mathbf{u}|\mathcal{X})\frac{\phi(\mathbf{u})}{\phi(\mathbf{u})}\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u} \\ &\geq \iint p(\mathbf{f}|\mathbf{u},\mathcal{X})\phi(\mathbf{u})\log\frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u}|\mathcal{X})}{\phi(\mathbf{u})}\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u}. \end{aligned} \tag{3.12}$$

Replacing $p(\mathbf{y}|\mathbf{f})$ with the site approximations from EP-DTC, in (3.12), leads to

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &\gtrsim \iint p(\mathbf{f}|\mathbf{u},\mathcal{X})\phi(\mathbf{u})\log\frac{\tilde{\mathbf{Z}}\mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})p(\mathbf{u}|\mathcal{X})}{\phi(\mathbf{u})}\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u} \\ &\gtrsim \int \phi(\mathbf{u})\left(H + \log\frac{\tilde{\mathbf{Z}}p(\mathbf{u}|\mathcal{X})}{\phi(\mathbf{u})}\right)\mathrm{d}\mathbf{u}, \end{aligned} \tag{3.13}$$

where

$$H = \int p(\mathbf{f}|\mathbf{u},\mathcal{X})\log\tilde{\mathbf{Z}}\mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})\mathrm{d}\mathbf{f}. \tag{3.14}$$

Now, let $\boldsymbol{\alpha} = \mathbf{K}_{\mathbf{fu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u}$. H can be re-expressed as

$$\begin{aligned} H &= -\frac{n}{2}\log 2\pi - \frac{1}{2}|\tilde{\boldsymbol{\Sigma}}| - \int p(\mathbf{f}|\mathbf{u},\mathcal{X})(\mathbf{f}-\tilde{\boldsymbol{\mu}})^\top\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{f}-\tilde{\boldsymbol{\mu}})\mathrm{d}\mathbf{f} + \sum_{i=1}^{n}\log\tilde{Z}_i \\ &= -\frac{N}{2}\log 2\pi - \frac{1}{2}|\tilde{\boldsymbol{\Sigma}}| - \frac{1}{2}\operatorname{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\tilde{\boldsymbol{\mu}}\boldsymbol{\alpha}^\top + \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top)\tilde{\boldsymbol{\Sigma}}^{-1}\right) \\ &\quad - \frac{1}{2}\operatorname{tr}\left((\mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}})\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \sum_{i=1}^{n}\log\tilde{Z}_i \\ &= \log\mathcal{N}(\tilde{\boldsymbol{\mu}}|\boldsymbol{\alpha},\tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}\operatorname{tr}\left((\mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}})\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \sum_{i=1}^{n}\log\tilde{Z}_i. \end{aligned} \tag{3.15}$$

Using (3.15) in (3.13), and reversing Jensen's inequality, in Equation (3.12), leads to the definition of the lower bound on $\log p(\mathbf{y}|\mathbf{X})$

$$\begin{aligned}\mathcal{L}_E &= \log \int \mathcal{N}(\tilde{\boldsymbol{\mu}}|\boldsymbol{\alpha}, \tilde{\boldsymbol{\Sigma}}) p(\mathbf{u}|\mathbf{X}) \mathrm{d}\mathbf{u} - \frac{1}{2}\operatorname{tr}\left((\mathbf{K_{ff}} - \mathbf{Q_{ff}})\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \tilde{\mathbf{Z}} \\ &= \log \mathcal{N}(\tilde{\boldsymbol{\mu}}|\mathbf{0}, \mathbf{Q_{ff}} + \tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}\operatorname{tr}\left((\mathbf{K_{ff}} - \mathbf{Q_{ff}})\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \tilde{\mathbf{Z}} \\ &\lesssim \log p(\mathbf{y}|\mathbf{X}).\end{aligned} \tag{3.16}$$

Notice that $\mathcal{L}_E$ has the same shape of the lower bound $\mathcal{L}_T$ (equation (3.1)), only rather than being weighted by the noise variance from the process, the elements of the trace are now weighted by the variances from the site approximations. The trace term in Equation (3.16) forces $\mathbf{Q_{ff}}$ to be closer to $\mathbf{K_{ff}}$, preventing overfitting and also adding more stability to the computations.

We now compare the quality of the new bound $\mathcal{L}_E$ with EP-FITC. We applied both approximations to a set of classification benchmarks (12 data sets: two from Ripley's collection[3] and 10 from Gunnar Rätsch's benchmarks[4]). Table 3.1 shows the error and negative log-probabilities obtained with each model. The number of inducing inputs used was the same for both models in each case. The covariance functions were all taken to be an exponentiated quadratic or RBF with white noise. The values in the table correspond to the average results of 10 folds over the data (except for the synthetic data set, which is already divided into test and training sets). In the case of the *crabs* data set, we randomly created 10 test/train partitions of size 80/120 ensuring that each training set had equal number of observations per class. Rätsch's benchmark contains 100 training and test splits per data set. In these experiments, we worked with 10 splits randomly chosen. Hyperparameters and inducing inputs were optimized jointly by scale conjugate gradients. For each split, we tried three different initializations and retained the model with the highest marginal likelihood for testing. Both models exhibited a similar performance, with EP-FITC being marginally better.

3.4 Log-Gaussian Cox Process with EP

As we have mentioned, we are particularly interested in exploring ways of extending our modelling framework with the tools needed to handle different kind of spatiotemporal

[3] http://www.stats.ox.ac.uk/pub/PRNN/.
[4] http://theoval.cmp.uea.ac.uk/~gcc/matlab.

				EP-FITC		**Var EP-DTC**	
data set	q	m	train/test	error	nlp	error	nlp
synthetic	2	4	250/1000	0.0910	*0.2595*	0.0930	*0.2618*
crabs	5	10	80/120	0.0450	*0.2493*	0.0458	*0.2943*
banana	2	20	400/4900	0.1092	*0.2535*	0.1083	*0.2543*
breast-cancer	9	2	200/77	0.2610	*0.5242*	0.2805	*0.5363*
diabetes	8	2	468/300	0.2273	*0.4789*	0.2290	*0.4922*
flare-solar	9	3	666/400	0.3410	*0.5932*	0.3250	*0.5959*
german	20	4	700/300	0.2470	*0.4985*	0.2637	*0.5114*
heart	13	2	170/100	0.1600	*0.4003*	0.1610	*0.4221*
thyroid	5	6	140/75	0.0560	*0.2087*	0.0560	*0.2164*
titanic	3	2	150/2051	0.2373	*0.5180*	0.2368	*0.5274*
two-norm	20	2	400/7000	0.0239	*0.1273*	0.0241	*0.1682*
waveform	21	10	400/4600	0.0966	*0.2406*	0.0995	*0.2682*

Table 3.1 Sparse binary classification models comparison. EP-FITC approximation is compared with variational EP-DTC across different data sets. For each data set, columns q and m show the input dimensionality and the number of inducing inputs used. Column train/test shows the number of instances in each of the training and test sets. The error corresponds to the ratio of misclassified instances in the test set. Column nlp shows the negative log-probability of the test instances.

processes. Due to its relevance in the field of spatial statistics, we present here an example of a Poisson process fitted using the variational EP-DTC approximation.

Consider an *inhomogeneous Poisson process* (see Appendix C.2) on a domain $\mathbb{S}$, parametrized by a rate $\lambda : \mathbb{S} \to \mathbb{R}^+$. The number of events within a region $\mathbb{B} \subset \mathbb{S}$ has a Poisson distribution with parameter $\lambda_{\mathbb{B}} = \int_{\mathbb{B}} \lambda(s)\mathrm{d}s$. Once we know the shape of $\lambda(\cdot)$ across the space $\mathbb{S}$, we can characterize the whole process. We cannot model $\lambda(\cdot)$ as a GP due to the restriction of it being positive. However, we can define it as a transformation over $(f_s) \sim \mathcal{GP}$. For example, if $f : \mathbb{S} \to \mathbb{R}$, then $\exp(f) : \mathbb{S} \to \mathbb{R}^+$. This transformation is what makes Bayesian inference intractable, and it is precisely why we resort to an approximation. We generated a toy data set of 130 points that represents the observations of a count process. In Figure 3.1, we compare the model fit of a GP regression model (Gaussian noise assumption) and Poisson model using the standard EP algorithm, and a Poisson process using the variational EP-DTC approximation with 7 inducing inputs. A clear flaw of the GP regression model (top image) is that its predictions allow the process to be negative, which is not consistent with the process that generated the data. The Poisson model provides a more realistic

behaviour of the predictions. The model fit is not very different when using EP (central image) or the variational EP-DTC algorithms. The former presents closer predictions to the data points, but the later can be improved by increasing the number of inducing inputs.

Table 3.2 compares the fit of these three models to the data, based on a 5-fold cross-validation (see Appendix D for more details on cross-validation), and the execution time required by each one. The results show that, provided we can ignore the positiveness and non-continuity of the data, the regression model has the highest predictive probabilities[5], followed by the EP model and at last the variational EP-DTC model. In applications where either discreteness or positiveness cannot be disregarded, the comparison of the regression model using predictive probabilities is not valid. If the assumptions of the model ignore what is known about the data, the resulting predictive probabilities would lack of meaning.

Execution times are not a minor factor to consider, specially when the differences between the three methods are in terms of orders of magnitude[6]. Although the EP approximation provides the most accurate model, considering model assumptions and CV score, it is also the slowest of the three algorithms. Var EP-DTC is computed much faster than EP, at the expense of lower predictive probabilities. The fastest method, due to its analytical solution, is the regression model. The last makes, once more, the regression model an appealing alternative for using, if possible. In the light of the execution cost and the data fit, we have to ponder the importance of our estimates being discrete and positive. We have to ponder how much we gain or lose in terms model and predictions accuracy. For instance, it is known that $\text{Poisson}(\lambda) \to \mathcal{N}(\lambda, \lambda)$, when $\lambda \to \infty$. Which means that the Gaussian assumption, and therefore the regression model, can be considered as valid when modelling large counts[7].

[5]Since the regression model uses a continuous observation model, its predictive probabilities were computed as $F(y_i + .5|\mathbf{y}_{-i}, \mathbf{X}) - F(y_i - .5|\mathbf{y}_{-i}, \mathbf{X})$, where F is the corresponding cumulative distribution function.

[6]The execution times reported are not an absolute measure. There might be faster implementations of the algorithms than the ones we used.

[7]Also notice that the farther the counts are from zero, the smallest the probability assigned to negative outputs in a regression model. In this scenario, the positiveness assumption can be considered as satisfied.

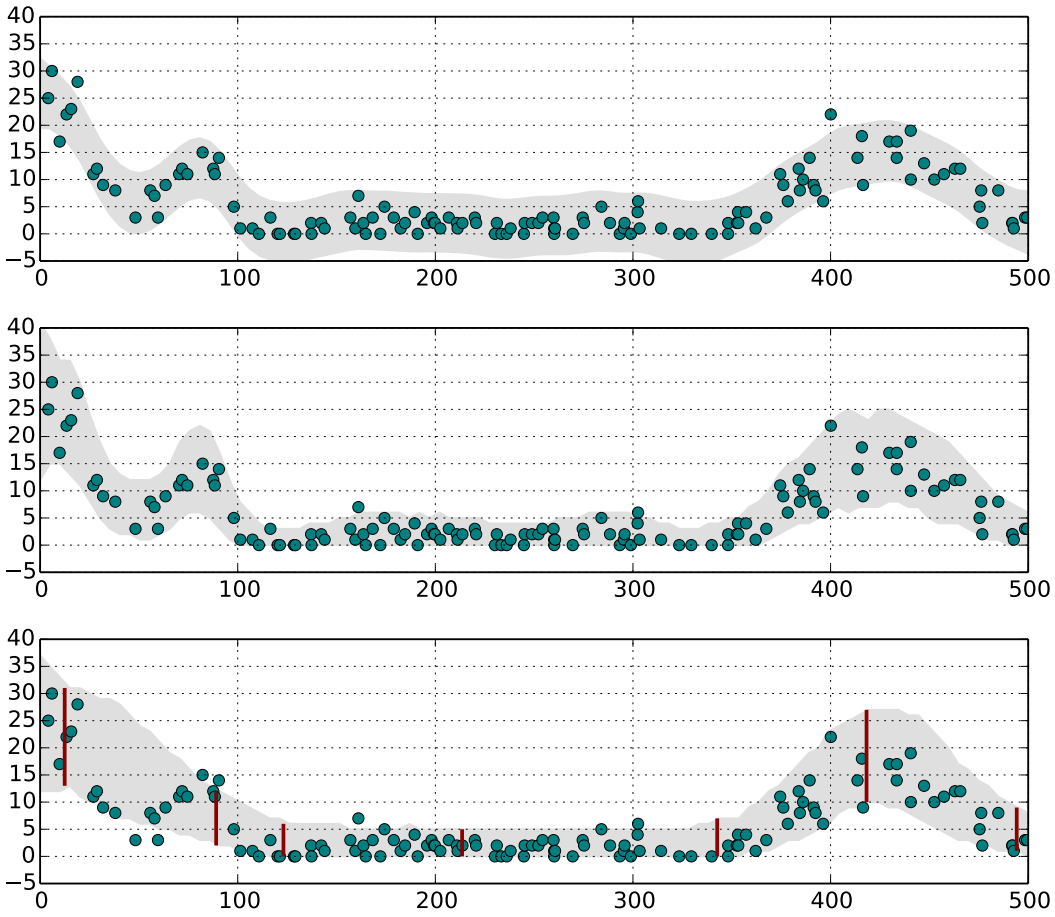


Fig. 3.1 Poisson regression with EP. Top panel shows the model fit using standard GP regression model. Middle panel shows a Poisson model using EP approximation. Bottom panel shows a Poisson model using variational EP-DTC. The dots show the observed data. The grey area represents the 95% credibility interval and the vertical lines indicate the optimized inducing input locations.

Model	CV	time
Regression	-255.16	0.2274
EP	-301.96	318.6685
Var EP-DTC	-331.68	64.2066

Table 3.2 Algorithms for modelling count data (regression, EP and Var EP-DTC). For each algorithm, the table shows the 5-fold cross-validation results (CV) and the execution time measured in seconds.

3.5 Final Comments

We have introduced a sparse approximate inference method able to handle non-Gaussian data. In the comparisons presented, our approximation has similar results to the EP-FITC approximation. The potential of this model has not been fully explored so far. As we will show in the next chapter, this model can be used to assimilate high-dimensional data with different noise sources and analyze it with dimensionality reduction techniques.

We will start by showing how it is possible to extend the variational bound in Equation (3.16) to handle uncertainty on the inputs of the Gaussian process. Girard et al. (2003) and Girard and Murray-Smith (2005) are able to work with noisy inputs in the *predictions* of a GP regression model, by propagating the uncertainty through the covariance. We additionally use variational inference to approximate the *marginal likelihood* and incorporate uncertain inputs in the training procedure. This makes possible, within our framework, to handle uncertain inputs in classification models and to construct hybrid continuous-discrete dimensionality reduction models.

Chapter 4

Hybrid Discriminative-Generative Approach

Urtasun and Darrell (2007) proposed a GP classification method that uses latent variable models trained with discriminative priors over the latent space. Their model uses a discriminative approach in the latent space, but preserves the generalization properties of a generative model. Inspired by this work, in this chapter, we work towards extending the Gaussian process classification to allow propagation of a generative model through the conditional distribution. This is achieved through a marriage of expectation propagation (Minka, 2001; Opper and Winther, 2000) with the variational approximations of Titsias (2009) and Titsias and Lawrence (2010). The resulting framework allows us to deal with mixed discrete-continuous data. We apply it to classification with missing and uncertain inputs, visualization of hybrid binary and continuous data and joint manifold modelling of labelled data.

Non-Gaussian data has already been considered in the context of continuous latent variables. The bound of Jaakkola and Jordan (1996) was applied to unsupervised learning by Tipping (1999) for the principal component analysis (PCA) of binary data (see also Lee and Sompolinsky (1999); Schein et al. (2003)). These models are related to GP models due to the shared challenge of combining a Gaussian prior with a non-Gaussian likelihood. This arises due to the duality between the latent variables (in this case, equivalent to the *inputs* $\mathbf{X}$) and desired principal subspace generated by the mapping $\mathbf{W} \in \mathbb{R}^{p\times q}$ in PCA. By associating the j-th column of the mapping matrix $\mathbf{w}_j$ with the j-th output dimension of the data $\mathbf{y}_j$, the associated mapping of the latent variables can be expressed as $\mathbf{y}_j = \mathbf{X}\mathbf{w}_j$. Factors $\mathbf{w}_j$ are induced to be jointly Gaussian distributed, as in a GP, by defining the usual spherical Gaussian prior independently over the latent variables $x_{ij} \sim \mathcal{N}(0, 1)$. Indeed, marginalizing $\mathbf{w}_j$ with

a Gaussian prior leads directly to a GP with a linear covariance function. This was the relation exploited by Lawrence (2005) to generalize PCA in the Gaussian process latent variable model (GP-LVM).

4.1 Discriminative and Generative Models

Discriminative models or regression models estimate a conditional density $p(\mathbf{y}|\mathbf{X})$, so that for any given $\mathbf{x}_*$ the probability of a new output y_* is known. *Generative models*, consist of estimating the joint distribution between response and *latent predictors* $p(\mathbf{y}, \mathbf{X})$. The use of the term latent predictors is due to the training of these models does not rely on a pairing of desired outputs and inputs. Dimensionality reduction techniques are an example of generative models, where for a given data set $\mathbf{Y} \in \mathbb{R}^{n \times p}$ a lower dimensional representation of latent variables $\mathbf{X} \in \mathbb{R}^{n \times q}$ (for $q < p$) is constructed.

Gaussian processes have been reformulated as a generative model known as the Gaussian process latent variable model (Lawrence, 2005). In this model, a GP provides a probabilistic mapping between $\mathbf{X}$ and $\mathbf{Y}$. As initial assumption, GP-LVM considers the dimensions of $\mathbf{Y}$ to be independent conditioned on the features. Then $p(\mathbf{Y}|\mathbf{X})$ can be written as

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{p} p(\mathbf{y}_j|\mathbf{X}), \tag{4.1}$$

where $\mathbf{y}_j$ represents the j-th column of $\mathbf{Y}$. The exact marginal likelihood of the data can then be computed as the expectation of discriminative models for each dimension; this is

$$\begin{aligned} p(\mathbf{Y}) &= \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X} \\ &= \prod_{j=1}^{p} \int p(\mathbf{y}_j|\mathbf{X}) p(\mathbf{X}) \mathrm{d}\mathbf{X} \\ &= \prod_{j=1}^{p} \langle p(\mathbf{y}_j|\mathbf{X}) \rangle_{p(\mathbf{X})}. \end{aligned} \tag{4.2}$$

Difficulty in computing such expectations arises from $\mathbf{X}$ being non-linear inside $p(\mathbf{y}_j|\mathbf{X})$. In the original paper, the latent variables $\mathbf{X}$ were optimized by maximum likelihood. Later, Titsias and Lawrence (2010) showed that they can be approximately marginalized through a collapsed variational approach (Hensman et al., 2012), analogue to the sparse variational approximation (see Section 2.4). They introduced the factorized

variational distribution

$$q(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_i, \mathbf{S}_i), \tag{4.3}$$

where $\{\mathbf{S}_i|i = 1, \ldots, n\}$ are diagonal matrices. A lower bound on $\log p(\mathbf{Y})$ is defined using (4.3), and $\phi(\cdot)$ is determined through a mean field approach. This allows the uncertainty in the latent space to be incorporated in the model and the underlying dimensionality to be determined. Damianou et al. (2012) developed the manifold relevance determination (MRD) as a means to determine the latent dimensionality in the context of multi-view learning. In their approach, latent variables are automatically allocated to the *relevant* views. As a result, some latent dimensions are shared across the views, whilst other are private to a particular one. So far, however, this model has only been applicable to Gaussian data. Here, we extend their approach to non-Gaussian data. The resulting framework allows a range of model extensions including:

1. Classification with uncertain inputs.
2. Dimensionality reduction of non-Gaussian data.
3. Joint modelling of binary labels alongside a data set to form a discriminative latent variable model.

4.2 Hybrid Model

Lasserre et al. (2006) present a general framework for discriminative training of generative models, that relies on a model formulation with an additional set of parameters[1]. We follow a similar approach, by using a variational formulation. So far, we have assumed that we are given a full set of input-output pairs for each data point $(\mathbf{x}_i, y_i)$. The advantage of extending the variational formulation with EP is that we can now consider distributions over $\mathbf{x}_i$, which allows inference with uncertain inputs and multi-view learning for hybrid data sets. We will assume that we have a Gaussian approximation to the posterior density $q(\mathbf{X})$ (equation (4.3)) in place of $\mathbf{X}$.

Following Titsias and Lawrence (2010) and putting together Equations (4.2) and (4.3), leads to the lower bound

$$\log p(\mathbf{Y}) \geq \sum_{j=1}^{q} \langle \log p(\mathbf{y}_j|\mathbf{X}) \rangle_{q(\mathbf{X})} - \mathrm{KL}\left(q(\mathbf{X}) \| p(\mathbf{X})\right). \tag{4.4}$$

[1] Additional to the parameters of the discriminative and generative models.

Given an EP approximation, as in Equation (3.4), and following a procedure similar to the one in Section 3.3 (equations (3.12) to (3.16)), the following inequality arises

$$\begin{aligned}\langle \log p(\mathbf{y}_j|\mathbf{X})\rangle_{q(\mathbf{X})} \gtrsim \langle \log\langle(\mathcal{N}(\tilde{\boldsymbol{\mu}}_j|\boldsymbol{\alpha}_j, \tilde{\boldsymbol{\Sigma}}_j))\rangle_{p(\mathbf{u}_j|\mathbf{X})}\rangle_{q(\mathbf{X})} \\ - \frac{1}{2}\operatorname{tr}\left(\langle \mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}}\rangle_{q(\mathbf{X})}\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \log \tilde{\mathbf{Z}}.\end{aligned} \tag{4.5}$$

It follows that a lower bound on the log-marginal likelihood can be computed as

$$\begin{aligned}\mathcal{L}_H &= \log \mathcal{N}\left(\tilde{\boldsymbol{\mu}}|\mathbf{0}, \boldsymbol{\Psi}_1^\top \mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Psi}_1 + \boldsymbol{\Lambda} + \tilde{\boldsymbol{\Sigma}}\right) - \tilde{\psi}_0 \\ &\quad + \operatorname{tr}\left(\mathbf{K}_{\mathbf{uu}}^{-1}\tilde{\boldsymbol{\Psi}}_2\right) - \mathrm{KL}\left(q(\mathbf{X})\|p(\mathbf{X})\right) + \log \tilde{\mathbf{Z}} \\ &\lesssim \log p(\mathbf{Y}),\end{aligned} \tag{4.6}$$

where $\tilde{\psi}_0 = \operatorname{tr}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\langle \mathbf{K}_{\mathbf{ff}}\rangle_{q(\mathbf{X})}\right)$, $\boldsymbol{\Psi}_1 = \langle \mathbf{K}_{\mathbf{uf}}\rangle_{q(\mathbf{X})}$, $\tilde{\boldsymbol{\Psi}}_2 = \langle \mathbf{K}_{\mathbf{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K}_{\mathbf{fu}}\rangle_{q(\mathbf{X})}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix such that $\boldsymbol{\Lambda}_{ii} = \operatorname{tr}\left(\tilde{\boldsymbol{\Psi}}_{2(i)}\mathbf{K}_{\mathbf{uu}}^{-1}\right) - \boldsymbol{\Psi}_{1(i)}^\top \mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Psi}_{1(i)}$. The subindex (i) means that we are only taking the i-th column of the corresponding matrix.

Notice that $\mathcal{L}_H$ has no longer the form of the DTC approximation. Instead, its form is closer to the FITC approximation[2], as it uses a covariance matrix that can be expressed as the sum of a diagonal and a non-diagonal matrices. An EP algorithm can be implemented for this new covariance form. Updates computation in this new algorithm resemble those of EP-FITC (Naish-Guzman and Holden, 2008), but the origin of the terms in the covariance is conceptually different.

4.2.1 Structure of the Posterior Moments

Analogue to the formulation of EP-DTC, but using a covariance matrix as in Equation (4.6), we start with a prior covariance of the form $\mathcal{N}(\mathbf{0}, \hat{\boldsymbol{\Psi}}^\top \mathbf{R}^\top \mathbf{R}\hat{\boldsymbol{\Psi}} + \hat{\mathbf{L}})$, where $\hat{\mathbf{L}} \in \mathbb{R}^{n\times n}$ is a diagonal matrix, $\mathbf{R}$ is the Cholesky decomposition of $\mathbf{K}_{\mathbf{uu}}^{-1}$ and $\hat{\boldsymbol{\Psi}} \in \mathbb{R}^{m\times n}$. From the combination of this prior with an EP posterior approximation $\mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, we obtain a posterior mean and covariance of the form

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right), \tag{4.7}$$

$$\boldsymbol{\Sigma} = \left(\tilde{\boldsymbol{\Sigma}}^{-1} + \left(\hat{\boldsymbol{\Psi}}^\top \mathbf{R}^\top \mathbf{R}\hat{\boldsymbol{\Psi}} + \hat{\mathbf{L}}\right)^{-1}\right)^{-1}. \tag{4.8}$$

[2]The marginal likelihood in the FITC approximation is given by $\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{Q}_{\mathbf{ff}} + \operatorname{diag}\left(\mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}}\right) + \sigma^2\mathbf{I})$.

As we will see next, the covariance structure in the prior will be kept in the posterior. By applying the matrix inversion lemma to Equation (4.8), $\boldsymbol{\Sigma}$ can be re-expressed as

$$\boldsymbol{\Sigma} = \left(\tilde{\boldsymbol{\Sigma}}^{-1} + \hat{\mathbf{L}}^{-1} - \hat{\mathbf{L}}^{-1}\hat{\boldsymbol{\Psi}}\mathbf{R}^{\top}\left(\mathbf{R}\hat{\boldsymbol{\Psi}}^{\top}\hat{\mathbf{L}}^{-1}\hat{\boldsymbol{\Psi}}\mathbf{R}^{\top} + \mathbf{I}\right)^{-1}\mathbf{R}\hat{\boldsymbol{\Psi}}^{\top}\hat{\mathbf{L}}^{-1}\right)^{-1}. \tag{4.9}$$

After applying a second time the matrix inversion lemma, to get rid of the negative exponent in Equation (4.9), we find that

$$\boldsymbol{\Sigma} = {}_{u}\hat{\boldsymbol{\Psi}}\ {}_{u}\mathbf{R}^{\top}\ {}_{u}\mathbf{R}\ {}_{u}\hat{\boldsymbol{\Psi}}^{\top} + {}_{u}\hat{\mathbf{L}}^{\top}, \tag{4.10}$$

for some suitable ${}_{u}\hat{\boldsymbol{\Psi}}$, ${}_{u}\mathbf{R}$ and ${}_{u}\hat{\mathbf{L}}$.

Due to the structure of the covariance matrix, the posterior mean $\boldsymbol{\mu}$ will also preserve a structure of the form

$$\boldsymbol{\mu} = \boldsymbol{\omega} + \hat{\boldsymbol{\Psi}}\boldsymbol{\gamma}, \tag{4.11}$$

for some $\boldsymbol{\omega} \in \mathbb{R}^n$ and $\boldsymbol{\gamma} \in \mathbb{R}^m$. Notice that substituting (4.10) into (4.7), leads to

$$\boldsymbol{\mu} = {}_{u}\boldsymbol{\omega} + {}_{u}\hat{\boldsymbol{\Psi}}\ {}_{u}\boldsymbol{\gamma}, \tag{4.12}$$

for some vectors ${}_{u}\boldsymbol{\omega}$ and ${}_{u}\boldsymbol{\gamma}$.

4.2.2 Update Computations

Equations (4.7) to (4.12), show the structure of the posterior mean and covariance. Based on these structures, we will now explain the low-rank update computations when changing an EP site approximation. Let the posterior mean and covariance be given by

$$\boldsymbol{\mu} = \boldsymbol{\omega} + \hat{\boldsymbol{\Psi}}\boldsymbol{\gamma}, \tag{4.13}$$

$$\boldsymbol{\Sigma} = \hat{\boldsymbol{\Psi}}\mathbf{R}^{\top}\mathbf{R}\hat{\boldsymbol{\Psi}}^{\top} + \hat{\mathbf{L}}, \tag{4.14}$$

and suppose that at the i-th iteration the natural parameters of the likelihood approximation are increased by $\Delta\tilde{\nu}_i$ and $\Delta\tilde{\tau}_i$. Then, the new posterior covariance and posterior mean can be computed by updating each one of their components as follows

$${}_{u}\hat{\mathbf{L}} = \hat{\mathbf{L}} - \frac{\Delta\tilde{\tau}_i\hat{\lambda}_{ii}^2}{1 + \Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i\mathbf{e}_i^{\top}, \tag{4.15}$$

$$ {}_u\hat{\mathbf{\Psi}} = \hat{\mathbf{\Psi}} - \frac{\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i\hat{\boldsymbol{\psi}}_i, \tag{4.16} $$

$$ \delta_i = \frac{\Delta\tilde{\tau}_i}{1+\Delta\tilde{\tau}_i s_{ii}}, \tag{4.17} $$

$$ {}_u\mathbf{R} = \text{chol}\left(\mathbf{R}^\top\left(\mathbf{I} - \mathbf{R}\hat{\boldsymbol{\psi}}_i\delta_i\hat{\boldsymbol{\psi}}_i^\top\mathbf{R}^\top\right)\mathbf{R}\right), \tag{4.18} $$

$$ {}_u\boldsymbol{\omega} = \boldsymbol{\omega} + \frac{(\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\omega_i)\hat{\lambda}_{ii}}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i, \tag{4.19} $$

$$ {}_u\boldsymbol{\gamma} = {}_u\hat{\mathbf{\Psi}}\boldsymbol{\gamma} + {}_u\hat{\mathbf{\Psi}}\left((\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\tilde{\mu}_i)\ {}_u\mathbf{R}^\top\ {}_u\mathbf{R}\ {}_u\hat{\boldsymbol{\psi}}_i\right), \tag{4.20} $$

where $\hat{\mathbf{L}} = [\hat{\lambda}_{ii}]$, $\hat{\boldsymbol{\psi}}_i$ is the i-th column of $\hat{\mathbf{\Psi}}$ and $\mathbf{e}_i$ is the i-th canonical basis vector of $\mathbb{R}^n$.

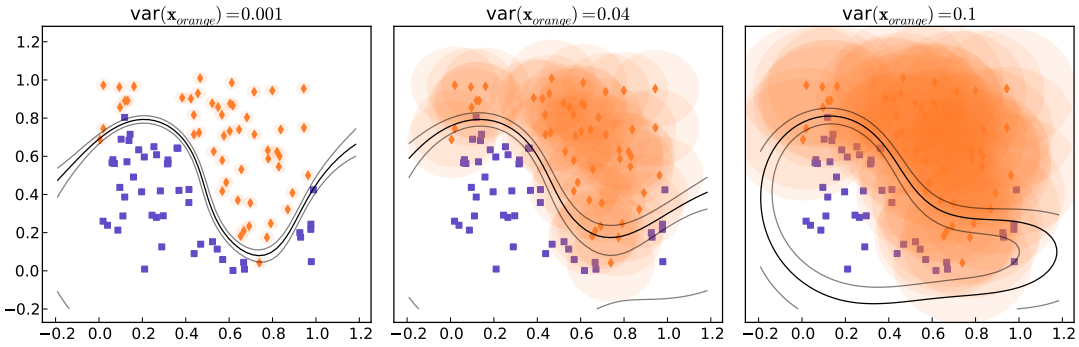
In the next sections, we consider applications of our model in three different domains: classification with uncertain inputs, dimensionality reduction of non-Gaussian data and classification using a hybrid discriminative-generative approach.

4.3 Classification With Uncertain Inputs

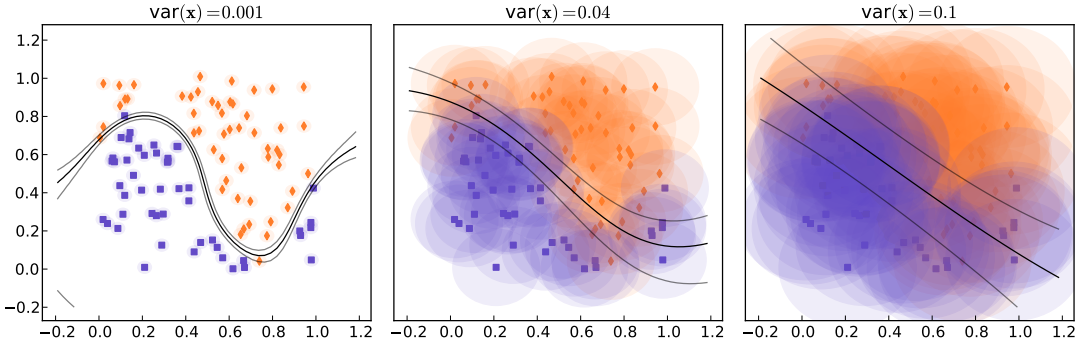
In probabilistic classification, we are not only interested in the class estimates, but also in a measure of the uncertainty about our predictions. If we are aware that there is uncertainty associated to the inputs on which the classification is based, it makes sense to incorporate this uncertainty in our predictions. Even if the class predictions do not change, credibility intervals may. In this section we present a couple of examples to illustrate how our framework handles such uncertainty.

4.3.1 Toy Example

In Figure 4.1, we show how the decision boundary in a classification model is affected by the increase in the inputs uncertainty. We considered an artificial binary classification problem. For an asymmetric increase in the uncertainty (Figure 4.1a), where only the inputs of one class become more uncertain, the decision boundary becomes more tightly wrapped around the inputs with less uncertainty. In contrast, when uncertainty



(a) Asymmetric uncertainty. The uncertainty increase on the inputs of one class only, from left to right, causes the decision boundary to shrink around the class with less uncertainty.



(b) Symmetric uncertainty. The uncertainty increase on the inputs of both classes, from left to right, causes a smoothing out of the decision surface.

Fig. 4.1 Classification with uncertain inputs. Class elements are distinguished by color and marker shape. The shaded ellipses represent 95% credibility intervals for each uncertain input. The contour lines represent the probabilities (bold line 0.5, light lines 0.4 and 0.6) of the points belonging to the *orange class*.

increases in both sets of input variables (Figure 4.1b) the decision boundary becomes much smoother overall.

4.3.2 Olivetti Face Data Set

Suppose we have a trained classifier for which the test point $\mathbf{x}_*$ has missing components. A simple solution would be to replace the missing values with the corresponding means from the training data. Our framework allows us to extend this idea by replacing the missing data with a Gaussian distribution, whose mean and variance matches the

	Without uncertainty		**With uncertainty**	
	error	nlp	error	nlp
No missing data	0.0200	*21.0248*		
50% of pixels randomly missing	0.1650	*94.8951*	0.1650	*73.5056*
Half of face occluded	0.1650	*69.1357*	0.1650	*67.0423*

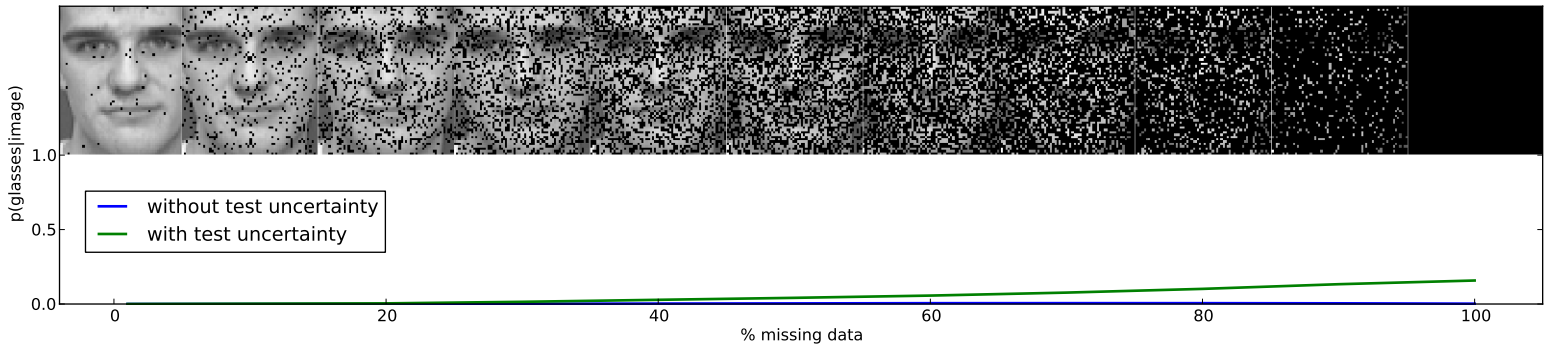
Table 4.1 Olivetti faces classification. Two models, one without input uncertainty and one with input uncertainty, are compared. The decision boundary is defined at a mean probability estimate of 0.5. The classification error is the same in both models, which means that, in this example, there is no difference in the decisions made. What changes is the negative log-probability (nlp). When input uncertainty is acknowledged, confidence in the estimates are reduced.

training data. We applied this idea using the *Olivetti face* data set[3] to predict whether or not a person is wearing glasses. We took a random 50/50 split to train two models: a standard GP-EP and a hybrid discriminative-generative model. On the test data, to simulate missing values, we removed a varying portion of pixels from the images (Figure 4.2). We then computed the class probability estimates of both models. As the proportion of missing values increases, the hybrid model becomes less certain and begin to converge towards the prior probability of an individual wearing glasses (about 30%). In contrast, the standard model just becomes certain that the image is a face with no glasses. Table 4.1 shows a comparison of the errors and negative log-probabilities obtained after introducing uncertainty.

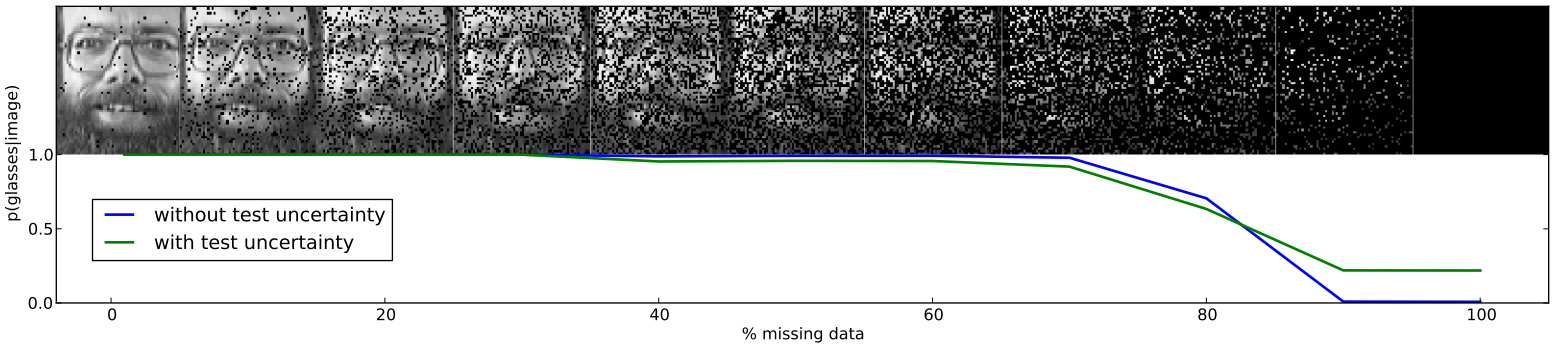
4.4 Dimensionality Reduction of Non-Gaussian Data

Manifold learning techniques model a high dimensional process, by encoding its dominant sources of variation in a latent process of lower dimensionality. Commonly, a Gaussian noise model is assumed, for example, in the probabilistic PCA and the Bayesian GP-LVM. By integrating EP to the GP variational framework, it is possible apply dimensionality reduction techniques on data with non-Gaussian noise. We applied our model on the *zoo* data set[4], where 101 animals from 7 categories (mammal, bird, fish, etc.) are described by 15 boolean attributes and 1 numerical attribute. The hybrid approach can model each attribute with a different noise model. We used a Bernoulli

[3] http://www.cs.nyu.edu/~roweis/data.html.
[4] http://archive.ics.uci.edu/ml/datasets/Zoo.



(a) Without glasses.



(b) With glasses.

Fig. 4.2 Classification with missing data. Increasing quantities of missing data are shown for two test cases, with the average (over 100 permutations) classification probability. For the standard GP-EP, missing pixels were replaced with the mean from the training data, for the hybrid model the independent marginal probability of the pixel is used. In the uncertain case, as more data are removed, the model predicts that the image contains glasses with $p = 0.3$, which matches the prior for the data set. Without consideration of the uncertainty, the model always predicts that the image contains glasses with probability 0, such is the appearance of the mean of the pixels.

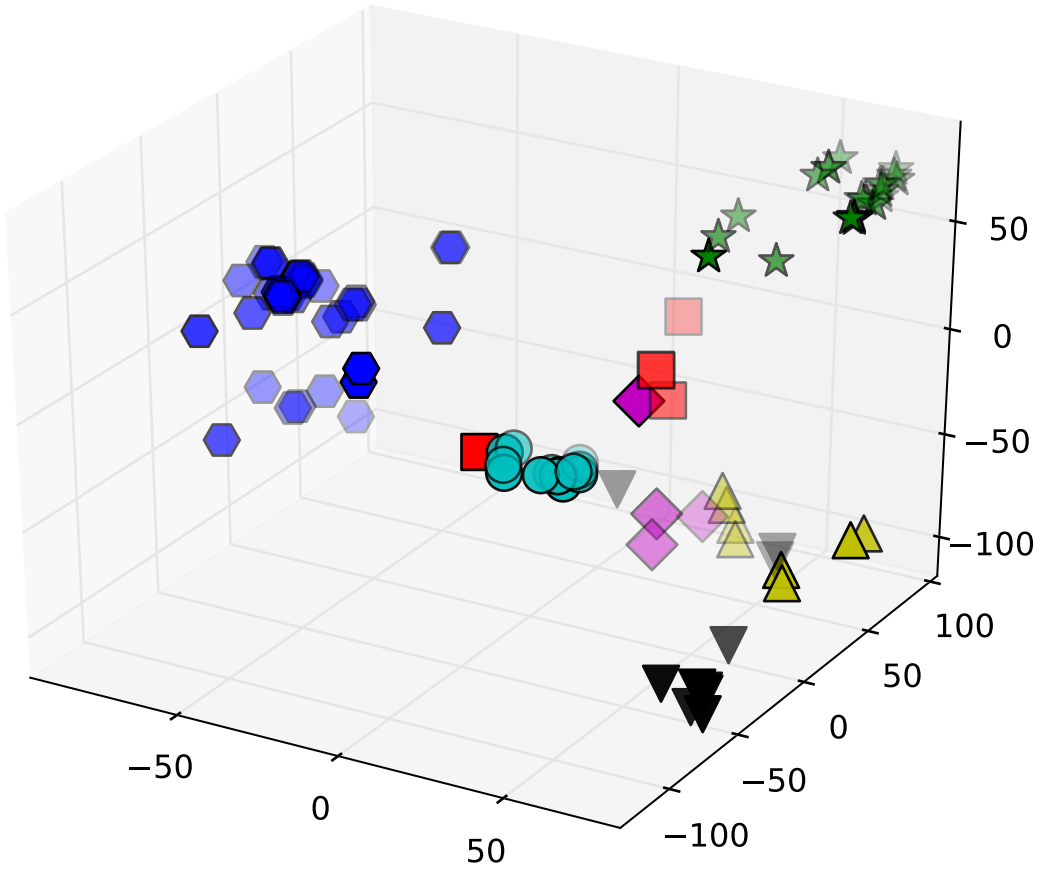


Fig. 4.3 Three dimensional representation of the *zoo* data set. The actual labels, unseen by the algorithm, are represented by different colors and bullets: mammals (blue hexagons), birds (green stars), reptiles (red squares), fish (cyan circles), amphibians (purple diamonds), insects (olive-green triangles) and crustaceans (black triangles).

and a Gaussian likelihoods for the boolean and numerical attributes, respectively. Figure 4.3 shows the latent representation of the data.

4.5 Discriminative Latent Variable Model

The manifold relevance determination approach of Damianou et al. (2012) considers multiple views of the same data set, allowing each view to be associated with private and shared portions of the latent space. We can construct a discriminative latent variable model, which includes class labels and data points as different views. We considered the 3s and 5s from the *USPS digits* database. In Figure 4.4, we show an

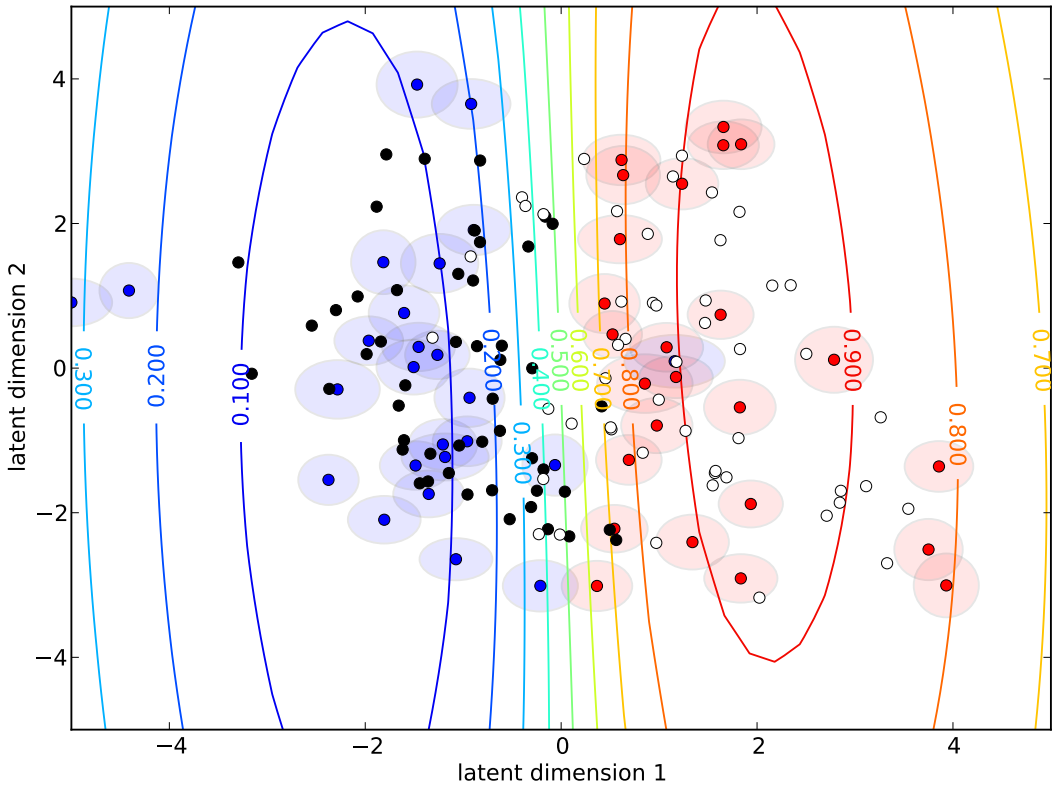
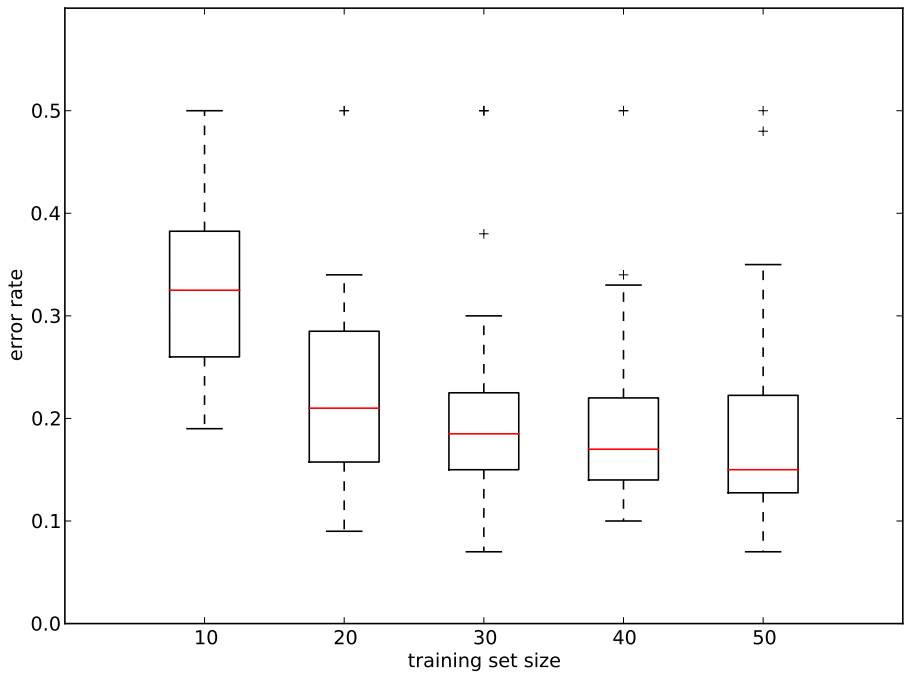


Fig. 4.4 Lower dimensional representation of the *USPS digits*. The blue and red points represent the examples of 3s and 5s, respectively, in the training set. The shaded ellipses represent the uncertainty of the latent variables. The black and white colors represent the test points (3s and 5s respectively) mapped to the learnt manifold. The contour lines represent the probability of an instance being five.

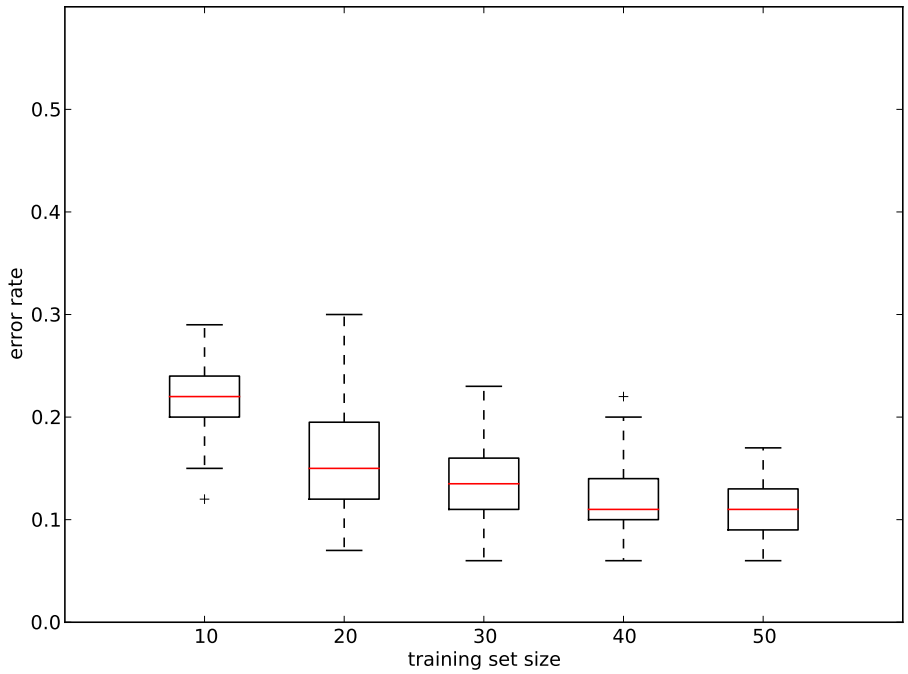
example where we used 50 observations to train the model and learn a 2-dimensional latent space. Notice that the discrimination occurs across the first latent dimension, whilst the second latent dimension is used to represent non-discriminative variation in the data. The figure shows the position of 100 unlabelled test data points mapped into the latent space alongside the locations of the training data.

4.6 Performance Against Generative Approach

We next followed Urtasun and Darrell (2007) in fitting a discriminative manifold model to labelled training sets of varying sizes. The error rates of the resulting models on 100 test points are shown in Figure 4.5a. Our results are similar to those presented by Urtasun and Darrell (2007). Our data set partitions differ and our error appears to



(a) Hybrid model.



(b) Standard EP-GP.

Fig. 4.5 Error rates as the data set size increases. Top panel shows the classification error rates on the *USPS digits* when using the hybrid model. Bottom panel shows the corresponding error rates when using standard EP-GP classification. Bar and whisker plots summarize 40 different subsampled training sets.

share the same form, but be worse overall. However, when we compared to standard EP-GP (Figure 4.5b), our performance was significantly worse. This contrasts to the results in Urtasun and Darrell (2007), who found standard GP classification underperforms on this data set. In our experience, standard EP-GP classification *can* perform badly when the initialization is poor and random restarts are not tried. This can explain the discrepancy between our results and theirs. To achieve similar results to EP-GP classification (and therefore exploit the advantages of the hybrid discriminative-generative model) we believe that our generative model needs to be more representative of the underlying data. One possible way in which this could be achieved would be through use of the deep GP formalism of Damianou and Lawrence (2013).

4.7 Final Comments

We have developed a framework for building hybrid discriminative-generative models with GP, by combining the EP approximation with a variational bound on the marginal likelihood. This required the development of a new sparse EP algorithm able to incorporate estimates of inputs uncertainty into the routine. These allowed us to incorporate discriminative Gaussian processes into a probabilistic model such as the Bayesian GP-LVM.

We have shown how the addition of inputs uncertainty leads to well behaved algorithms, in particular, when training on data where such uncertainty is class-dependent and when predicting using missing inputs. We are able to use these techniques to apply the Bayesian GP-LVM on non-Gaussian data and make continuous latent representations of mixed data types. The performance against generative approach is not as good as originally thought. More work in this area is needed to improve the lower dimensional representation of the data. However this is a line of research that diverts from our original one.

In the next chapter, we will revert to the traditional approach to dealing with the data. The chapter will be focused on using the techniques studied so far for modelling health facility records of malaria in Uganda.

Chapter 5

On the Challenges of Assimilating Data

So far, the discussion about the GP framework and the new methods introduced has been mainly theoretical. Different experiments have been presented, but their aim has been to show the properties of the methods proposed. The examples shown have been tailored for the techniques reviewed. The interest behind this line of research and the reason for using the modelling framework discussed has a practical goal: develop spatiotemporal modelling tools that help understanding malaria spreading across Uganda.

In this chapter, we apply the GP framework and the techniques reviewed previously to model data from the Health Management Information System (HMIS). We do not use data from *ad-hoc* surveys; but administrative records created with a different purpose than defining inputs for a statistical model. This brings the benefit of having plenty of data, but as we will see, this also brings some challenges that need to be faced when defining a probabilistic model.

We start the chapter with a brief introduction on malaria: how it is spread, why it is a disease that matters to some populations and why spatiotemporal models are an adequate tool for analyzing it. Then we discuss the data features used in this project and some of the challenges we face when using it. We will see how these challenges impose a turning point in the methodological approach used. We then define a model for the disease case-counts that assimilates the data characteristics we have observed. Finally, we try to improve our model performance by incorporating environmental data into the learning routine.

5.1 About Malaria

The interaction between human beings and malarial parasites is very old. The footprints of malaria can be observed in different moments in the history of civilization and it is likely that there are even some footprints in the history of our genetic evolution. Carter and Mendis (2002) discuss how some human populations might have evolved, preferring certain polymorphisms, depending on the resistance to some malaria effects. Although the moment in which this disease started infecting humans is still an open question (Escalante et al., 1995; Liu et al., 2010), evidence points out that human populations have been stalked by malaria, at least, since the dawn of agriculture (Joy et al., 2003).

It was not until the end of the XIX century when it was finally understood that it is a parasitic disease and that it is transmitted to humans by mosquitoes (Cox, 2010). The bite of the female Anopheles, seeking blood to complete its own reproductive cycle, is an essential step in the reproduction of the malarial parasite. It is due to this furtive vector and to the complex life cycle of the parasite that malaria has been such a burden and difficult to eradicate.

More than a century after discovering its transmission mechanism, malaria has been successfully eradicated from different regions of world (Trigg and Kondrachine, 1998). However, it is still endemic in 100 countries and represents a threat for 3.3 billion people approximately (World Health Organization and others, 2014). In Uganda, malaria is among the leading causes of morbidity and mortality (World Health Organization, 2015). Hospital data from 2010 and 2011 show that malaria was responsible for 22% of morbidity cases and 21% of deaths. The percentage of hospital deaths went up to 27% when considering only children under five (Ministry of Health, Health Systems 20/20, and Makerere University School of Public Health, 2012).

Different types of interventions can be carried on to prevent and treat malaria, such as vector or larva control, chemoprevention for vulnerable groups, or timely treatment (World Health Organization and others, 2014). The success of such interventions depend on how well the disease can be anticipated and how fast the population reacts to it. In this regard, mathematical modelling can be a strong ally for decision-making and health services planning.

5.2 Modelling When and Where

The life cycle of malarial parasites cannot be completed in regions that are not suitable for mosquitoes breeding. This makes malaria a geographic phenomenon where factors like altitude, temperature or lack of water are critical (Bailey, 1982). With this in mind, a wide range of mathematical models have been proposed in recent years for trying to unravel the dynamics between the environment and the biology of malaria (Smith et al., 2012). Reinerand et al. (2013) identify 388 different mechanistic models of mosquito-borne pathogen transmission published between 1970 and 2010. A common idea across these models has been the incorporation of temperature as a driving pattern of transmission, and in some cases, other details of mosquito and larval ecology. There has also been an interest in modelling the pathogen infection in host, by including concepts such as *super infection*[1] (Portugal et al., 2011) or immunity (Good and Doolan, 1999). Reinerand et al. (2013) considers spatial heterogeneity and temporal variation as an unrepresented theme in the literature.

Spatiotemporal modelling for mapping and prediction of infection dynamics is a challenging problem. First of all, because of the costs and difficulties of gathering data. Second, because of the challenges of developing a sound theoretical model that agrees with the data observed. Gaussian processes are a standard tool for the spatial analysis of disease risk (Gosoniu et al., 2006; Kleinschmidt et al., 2000; Quinn et al., 2011). This provides an elegant non-parametric method for using distance information to make estimates of risk across a spatial field. Recent advances in inducing variable approximations have made the Gaussian process framework more practical (Álvarez and Lawrence, 2011; Vanhatalo, 2006).

5.3 About HMIS Data

The main source of information used in this project are the malaria case-counts records provided by the Epidemiology and Surveillance Unit from the HMIS, in Uganda. The HMIS is a reporting tool of the country with the function of providing information regarding the Health Sector, as a means to support planning and decision-making. Information systems are crucial for the health authorities of any country. Its benefits will strongly depend on the quality of the data provided, as well as on the capacity of the decision makers of responding to what the system reports (World Health Organization, 2010, 2011).

[1] An infection that occurs on top of an earlier infection.

The HMIS was initiated in 1997, as a replacement of the former Health Information System (HIS). The HIS was designed in 1985 to capture data about some health services and morbidity for some diseases, however, it was later realized that it was leaving out important health-management information such as infrastructure, drug management and medical equipment availability, among others (Kintu et al., 2005).

A recent assessment from WHO, concludes that HIMS is producing data of good quality at a national level (World Health Organization and Uganda Ministry of Health, 2011). At a district level, it points out the following remarks:

- "Completeness of district reporting is poor in 9% of districts and completeness of health facility reporting is problematic for one-third of the districts;
- Accuracy of reporting is only partly adequate, with 18% of the district reports zero or missing, 7% of the districts having extreme outliers, and 9% of the districts having major differences between the annual total and the sum of the monthly reports;
- District population projections for the denominators in 2010/2011 are estimated to be off by more than one-third for 22% of districts."

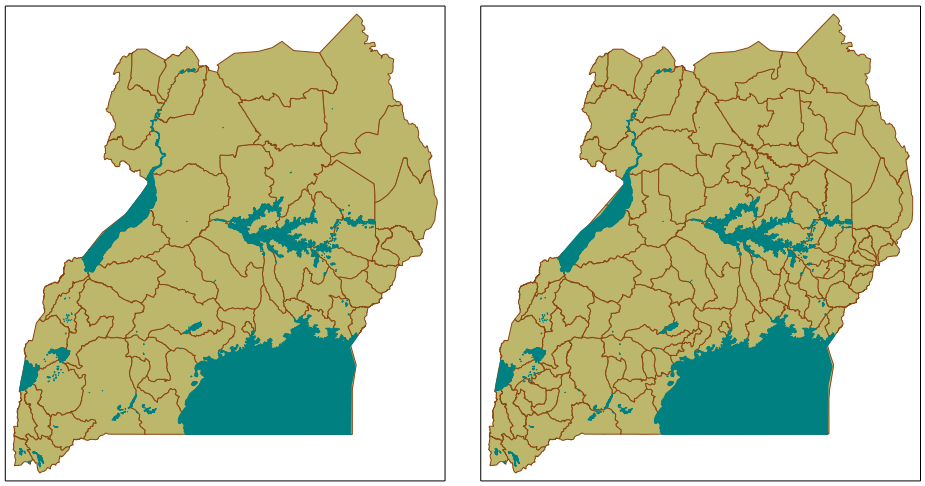
In addition to these remarks, through personal conversations with the health authorities, we found out that malaria records in HMIS correspond to people being treated for malaria, and not diagnosed with a specific test. Such tests are not part of routine mechanism, and therefore the records are very likely to be over-diagnosed by diseases with similar symptoms (Amexo et al., 2004; Castellani, 1907).

Due to the creation of new districts and changes in the boundaries of the existing ones, district-level data is not entirely consistent over this period. In 2003, there where only 56 districts, while today there are 112. Figure 5.1 shows a comparison between the district-limit definition in 2003 and today.

We had access to weekly information aggregated at a district level. No access to data at a hospital level or information specific to the individuals, such as age or gender, was possible. We analyzed data between January 2003 and July 2014.

5.4 Variation Sources in Malaria Records

In an ideal situation, with perfect health records, there would not be misreported cases of malaria or false positives/negatives. The only variation in the number of people reported to be infected would be originated by the actual evolution of the disease



(a) 2003 (b) 2015

Fig. 5.1 District boundaries in Uganda. The administrative delimitation of 2003 (left panel) consisted of 56 districts. Today's delimitation (right panel) consists of 112 districts.

case-counts. In practice, it is common to have records where variation comes from different sources, including changes in the reporting methodology or errors. Each of these sources has a particular way of disturbing the observed process. For example, it can be assumed that the natural variation in the case-counts of malaria is *smooth* across time; that changes in methodology affect in the form of bias[2]; and that errors are uncorrelated. Modelling all observed variation as part of the same source would not provide a good understanding of the phenomena. On the contrary, the conclusions drawn from such a model could just be misleading. An initial challenge is to understand the data and to identify where the main variation comes from.

As it was said before, we understand the part-properties by looking at the whole-properties and vice versa. Through the contrast between these two comes the identification of outliers and changes in the structure of the series. Figure 5.2 shows an example of the malaria records in Apac and Kotido districts. Marked with circles are observations that seem to behave in a different way than the rest. In the case of Apac, one could hypothesize that such outliers are in fact human errors, possibly an extra zero was added by accident. Kotido does not present observations with such extreme values, however there are a few that under some criterion can be thought of as being the result of a source of variation different from the infection process. HMIS

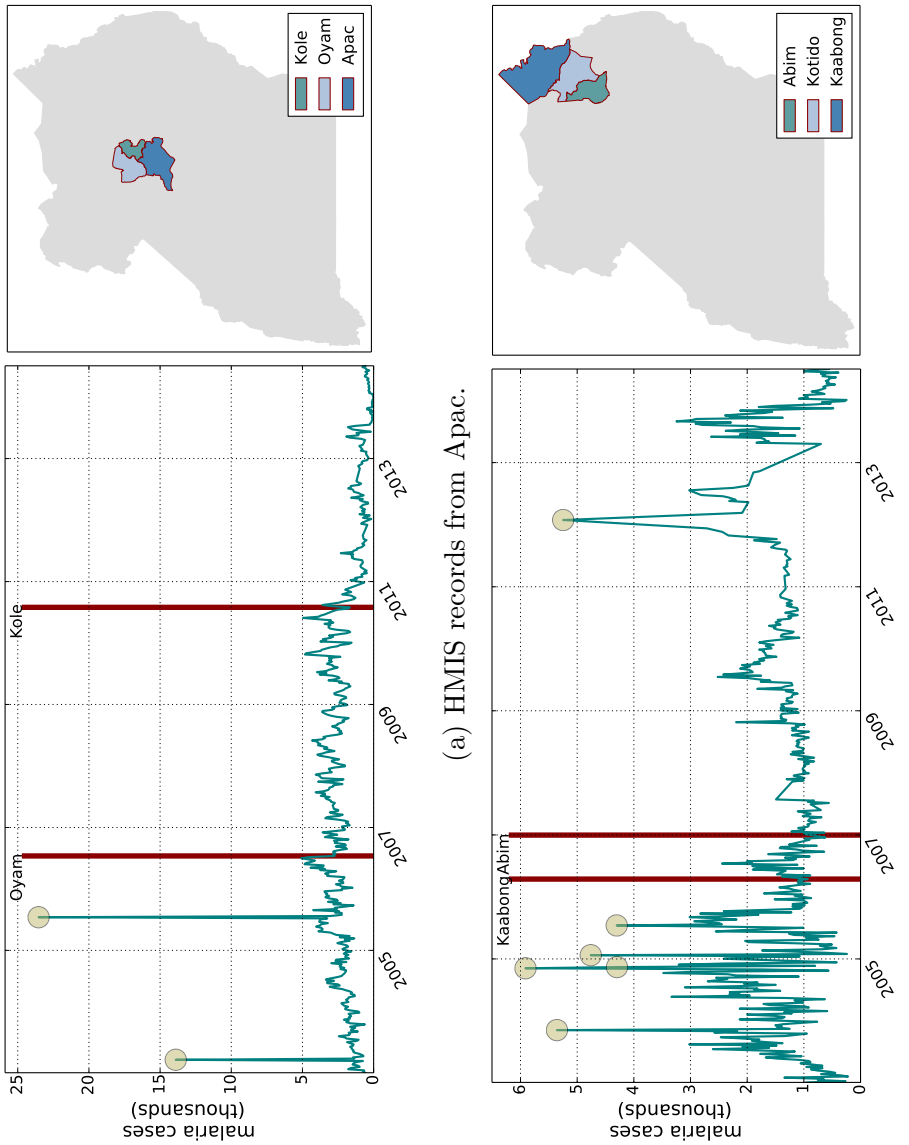
[2]For example, think of a bias induced due to a specific diagnostic procedure or due to a limited health service coverage in the population.

data is initially collected by hospitals or medical facilities, and then these records are aggregated per district. The creation of new districts by splitting the already existing ones has a clear effect in the structure of the series. The vertical red lines in the figures show the time points when the districts were split and part of its previous medical facilities started reporting independently as another district[3].

Another source of noise identified in the data is the inconsistency in the number of health facilities reporting. For each of the weekly counts of malaria per district, the HMIS database presents the total number facilities in the district and the number of reporting ones. Figure 5.3 shows a comparison of the malaria cases reported and the number of health facilities per observation in Arua and Iganga districts. The similarity of the trends, specially after 2009, indicates a strong effect of the misreporting facilities in the case-counts of malaria observed. This is, indeed, valuable information for understanding the database, unfortunately there is no information about the size of the facilities or their location that allows a more detailed inspection.

The analysis of the split districts (*parent* and *children*) aggregated provides some insight into the reporting process. Figure 5.4 shows the aggregated values of facilities and disease cases in Arua and Iganga, according to their definition in 2003. The cases of malaria registered have, in general, increased over the years (except when the number of reporting facilities is low). Before jumping to the conclusion of a worsen of the disease rate, it is important to notice the increment in the total number of facilities reporting. This suggest a larger base population where the disease is being diagnosed, which could mean that, rather than an increase of the infection, there is an improvement in the coverage of the health services. Another example of a significant change in the reporting process is exhibited in the left panel of figure 5.4a. Between 2007 and 2011, some of the facilities from Arua started reporting to Maracha, but it seems that later some of those started reporting back to Arua. This switch of reporting units is also evident in the disease case-counts reported in both districts. For example, see the drop-off of the disease trend in figure 5.3a. It is also interesting to notice that the effect of the district splitting is not eliminated by aggregating them. For example, Iganga presents an escalated increment in the number of facilities and in the malaria cases, after each splitting.

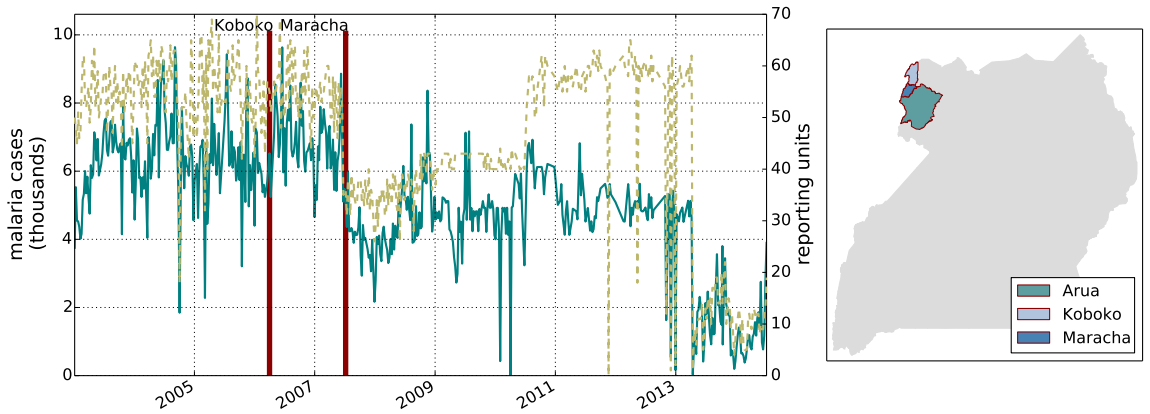
[3]This date does not correspond to the moment when the district was created, but just when its first records in the HMIS appeared. Before this point, it is not always clear when the records of the *parent* district stop containing records corresponding to the new district.



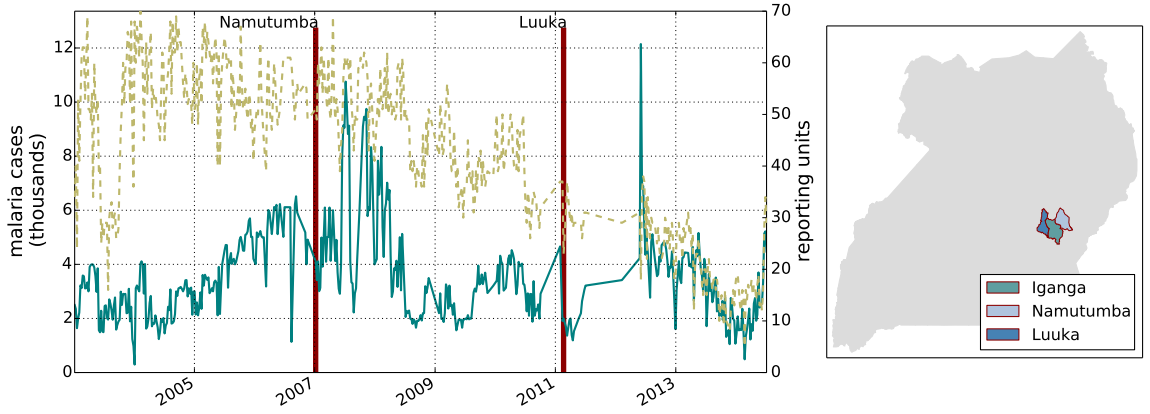
(a) HMIS records from Apac.

(b) HMIS records from Kotido.

Fig. 5.2 Atypical values in HMIS records. Circles show outliers which could be originated by errors in the records. Vertical red lines show the time point when a new district started reporting to the HMIS. Right panels show the original district (2003) and the current districts (2015) in which it has been split.

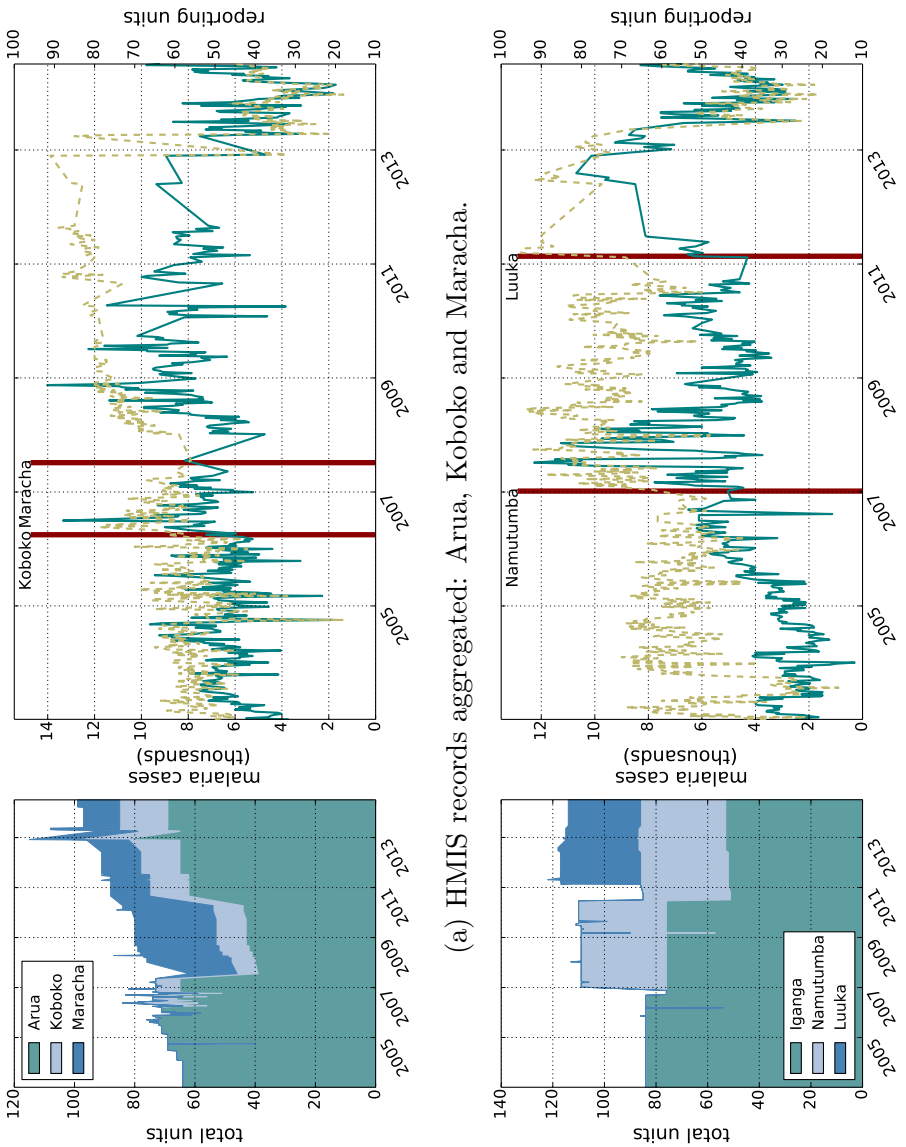


(a) HMIS records from Arua.



(b) HMIS records from Iganga.

Fig. 5.3 Malaria cases and health facilities reporting in HMIS records. Solid lines show the malaria cases and dashed lines show the number of reporting facilities. Vertical red lines show the time point when a new district started reporting to the HMIS. Right panels show the original district (2003) and the current districts (2015) in which it has been split.



(a) HMIS records aggregated: Arua, Koboko and Maracha.

(b) HMIS records aggregated: Iganga, Namutumba and Luuka.

Fig. 5.4 HMIS records aggregated. Left panels show the total number of health facilities per district, as it is registered each week. Right panels show the aggregated number of malaria cases (solid line) and the aggregated number of health facilities reporting (dashed line).

5.5 The Practical Limitations Imposed by the Data

The characteristics of the data analyzed have different implications on the modelling tools used. The aggregation at district level makes difficult the use of point patterns to study the malaria case-counts across space. The study of continuous spatial variation with the methods presented, in earlier chapters, is prevented by having data associated to very large polygons (districts). For this reason, our analysis will be focused only on the temporal variation of the number of malaria cases[4].

We have also discussed the costs associated to the implementation of approximate inference methods, in particular EP. In terms of training speed and predictive accuracy, we did not observe that EP-type methods outperform standard GP regression models. Considering, in addition, that we are interested in modelling 112 districts across many years, in this Chapter we find more convenient to put aside this technique. The different sources of variation and the several inconsistencies found in the data, forces us to think carefully about the noise model definition. Therefore, the approach we use will have to be different from what we have done previously.

5.6 A Model for HMIS Data

For doing inference on the HMIS data, we need to assimilate its different sources of variation. From the exploratory analysis of Section 5.4, we know that a model of HMIS data should be able to explain the disease case-counts in terms of its evolution in time and the number of facilities reporting. It should also be able to handle outliers or possible errors in the reporting process (see Figures 5.2a and 5.2b). The construction of such a model consists, first of all, of defining a sensible rule about how data observed is being corrupted from its original generator process. The last is a noise model that links the observations with the process described by the latent function. It is also needed to define a rule about how the different observations are related and the way in which variables like time or number of reporting facilities affect the disease case-counts.

An important factor to consider are the different boundaries definitions from 2003 until today. These changes might have an impact in the variation of the data observed (see Figures 5.3a and 5.3b). To simplify things, we will start analyzing the information of the districts within splitting periods, rather than across them. That means that we

[4]Markov Random Fields are commonly used when modelling data aggregated in polygons (Besag et al., 1991), however their use at this stage would deviate us from the framework we have been using so far.

will work with 168 virtual districts instead of 112. In a later stage of the analysis, we will work on a way to harmonize the data series.

5.6.1 Model Selection Criteria

Models are defined based on the characteristics of the process they represent. This way, some models are preferred over others simply because they represent better some assumptions about the data. Still it is common to end up with different choices, all of them reasonably valid. The task is then to decide which model is more convenient. When it is about choosing the parameters within a model, marginal likelihood optimization is among the preferred criteria. Selection between noise or covariance structures involves comparing models with different sets of parameters, and so marginal likelihood maximization is not a robust enough criterion for defining which model provides a better fit to the data. But it is not only a matter about fitting the observed data. The model should be able to provide insight about what has not been observed. Within the several ways of comparing and selecting models (see Vehtari et al. (2012) for a thorough review on the topic), out of sample cross-validation (CV) is a natural way of estimating the prediction error of a model (see Appendix D). Alternative methods are usually preferred to CV, as it involves the cost of re-fitting the model in the different training sets. However, in the case of leave-one-out cross-validation (LOO-CV) in GP regression models and EP approximations there are almost no computations required beyond the ones already carried on while fitting the model (Vehtari et al., 2014; Williams and Rasmussen, 2006). The cost is negligible. It is not the same situation for sparse approximations, as all the data points are encoded in the covariance matrix. It is not clear how to compute LOO-CV without re-fitting the model. When comparing sparse models, 5-fold cross-validation was used.

5.6.2 Noise Model Selection

The case-counts of malaria are a discrete and non-negative number. But given the large counts we are dealing with, we know that they can be safely represented with a Gaussian distribution. As we saw in Section 3.4, this turns out to be convenient as the log Gaussian Cox model involves a high execution cost versus a standard GP regression. An alternative, not discussed before, that ensures a positive distribution of the data is to assume a log-Gaussian distribution. Such a model can be worked out simply by using a Gaussian model on the logarithm of the data. In general, there could be a problem if there are observations with value zero, as the logarithm is only

defined for positive numbers. However, this is not the case here. We prefer to use this model over the Gaussian model to ensure positive estimates. For completeness, we present in Appendix E.1 a comparison of both models (Gaussian and log-Gaussian) using leave-one-out cross-validation. Given the difference in the data scales of each model (one is the transformed space of the other), to compare their predictive densities it is needed to rescale them with the Jacobian of the transformation (Gelman et al., 2014). We fitted models for data of 23 districts[5]. The predictive probabilities are not consistently higher for any of both models, and in a few cases they have similar values.

The model we will use for the HMIS data is given by

$$\log y_i = f_{\mathbf{x}_i} + \epsilon_i + \zeta_i, \tag{5.1}$$

where $(f_{\mathbf{x}_i}) \sim \mathcal{GP}$; $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is a noise term with homogeneous variance across observations; and ζ_i represents other sources of variation observed in the data and not explained by the previous terms (e.g., reporting errors).

The implications of working in the log-scale deserves special attention. The interaction of latent functions can be performed in the form of kernel additions or kernel products. The addition in the log-space represents a multiplicative effect in the original space. A multiplicative effect in the log-space, on the other hand, does not have a clear interpretation in the original space.

5.6.3 Kernel Selection

The structure dependence of the model is given by the kernel function used. By choosing to use a particular kernel, we are introducing into the model our beliefs about the process that generates the data. Our initial assumption about the infection process of malaria is that it evolves with some degree of smoothness across time. We need a kernel such that the closer the observations in time, the more similar values of the function it encodes. The Matérn kernel satisfies this condition, as it defines dependence through the distance between points with some exponential decay. It is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \frac{2^{1-\nu}\sigma^2}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu|\mathbf{x}_i - \mathbf{x}_j|}}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu|\mathbf{x}_i - \mathbf{x}_j|}}{\ell} \right), \tag{5.2}$$

where $\nu, \ell \in \mathbb{R}^+$ and K_ν is a modified Bessel function of order ν (Abramowitz and Stegun, 1972). The parameter ν controls the smoothness of the process, so that the

[5]We used a Matérn-3/2 kernel and only considered time as input.

larger the values of ν, the smoother the process. The formulation of the Matérn kernel in Equation (5.2) simplifies when ν is half integer. For example, the Matérn kernel with $\nu = 3/2$ (Mat-3/2) is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \left(1 + \frac{\sqrt{3|\mathbf{x}_i - \mathbf{x}_j|}}{\ell}\right) \exp\left(-\frac{\sqrt{3|\mathbf{x}_i - \mathbf{x}_j|}}{\ell}\right), \tag{5.3}$$

and the functions it describes are only once differentiable. In the limit, when $\nu \to \infty$, the Matérn kernel approaches the exponentiated quadratic kernel or RBF, which is given by

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\ell^2}\right), \tag{5.4}$$

and describes a function that is infinitely differentiable. Due to the noise of the HMIS data, a Mat-3/2 kernel seems more appropriate as initial assumption. Yet, we reckon any kernel in the Matérn family, including the exponentiated quadratic, would provide a sound model. In Appendix E.2.1 we show a comparison of the cross-validation results when using either a Mat-3/2 or RBF kernels. In general the Matérn kernel was preferred over the RBF kernel. For each district, the kernel with the highest LOO-CV predictive probabilities was used to model the data along this section of the chapter.

In addition to the smoothness across time, it seems sensible to think that the number of health facilities reporting has an effect on the case-counts of malaria observed. We also have seen some evidence of this above (Figure 5.3). We can define a linear relation between the variables using a kernel defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \alpha|\mathbf{x}_i - \mathbf{x}_j|^2, \tag{5.5}$$

for some $\alpha \in \mathbb{R}^+$.

Both kernels, the one to model the process across time and the one to model the dependence on health facilities, can be integrated as multiplicative effects. We define a multiplicative model by doing an orthogonal sum of both kernels[6].

Despite the evidence of the relation between the number of health facilities reporting and the number of malaria cases in some districts, we have to be aware that this might not be true for all districts in the country[7]. Even if there is a relation between these two

[6]Since each kernel works on a different dimension, the orthogonal sum yields a new kernel in two dimensions.

[7]There are many reasons why this can be the case. For example, the number of health facilities reporting might have errors; or the health facilities provide service to a very different number of patients so that the single number of facilities is not enough as explanatory variable.

variables, the linear kernel might not encode it properly. We have to decide whether adding the linear kernel benefits the model fit or not. Each district has its own specific characteristics and we are trying to be general enough to handle all of them. We compared a GP model using only a kernel to model time dependence with another that also considers a linear kernel on the number of health facilities reporting. We used LOO-CV to compare their predictive accuracy and choose the one with higher results. Appendix E.2.2 shows the results of this comparison. As an example, Figures 5.5 and 5.6 show the model fit to Kween and Ngora districts. In both districts, it is clear how the linear kernel helps correcting the effect of non-reporting facilities, which otherwise could be seen as a rapid drop in the malaria case-counts.

5.6.4 Outlier Detection

We have discussed in Section 5.4 that there might be some reporting errors in the HMIS database. To include these errors explicitly in Equation (5.1), we re-formulate it as

$$\log y_i = f_{\mathbf{x}_i} + \zeta_i + \epsilon_i, \tag{5.6}$$

where $f_{\mathbf{x}_i}$ is a function of time t_i and the number of reporting health facilities r_i, so that $(f_{\mathbf{x}_i}) \sim \mathcal{GP}$ with $\mathbf{x}_i = (t_i, r_i)$; $\zeta_i \sim \mathcal{N}(0, \sigma^2_{\zeta_i})$ is an error in the reporting process independent and with heterogeneous variance across observations; and $\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)$ is a noise term with homogeneous variance across observations that accounts for the residual variability in the model. At this stage we are still focused on modelling HMIS data between splitting points of the districts. Later on, when we try to harmonize the whole time series, a new term for dealing with this source of variation will be needed in Equation (5.6).

We expect reporting errors to occur only in some observations, and these being characterized by $\epsilon_i + \zeta_i >> \epsilon_i$ (see the observations marked with circles in Figure 5.2). If we assume that

$$\mathbf{z}_i = (\log y_i, r_i)^\top - (\log y_{i-1}, r_{i-1})^\top \sim \mathcal{N}(\dot{\boldsymbol{\mu}}, \dot{\boldsymbol{\Sigma}}), \tag{5.7}$$

for some $\dot{\boldsymbol{\mu}}$ and $\dot{\boldsymbol{\Sigma}}$, as the reporting errors are sparse, any point that contains a term $\zeta_i \neq 0$ will be unlikely under $\mathcal{N}(\dot{\boldsymbol{\mu}}, \dot{\boldsymbol{\Sigma}})$[8]. The parameters $\dot{\boldsymbol{\mu}}$ and $\dot{\boldsymbol{\Sigma}}$ can be learnt by Bayesian inference. The conjugate priors for this problem are an inverse-Whishart for

[8]We use a bivariate Gaussian that includes the reporting facilities because of the influence of this variable on the case-counts of malaria reported. An alternative would be to use a univariate Gaussian for $\tilde{z}_i = y_i/r_i - y_{i-1}/r_{i-1}$.

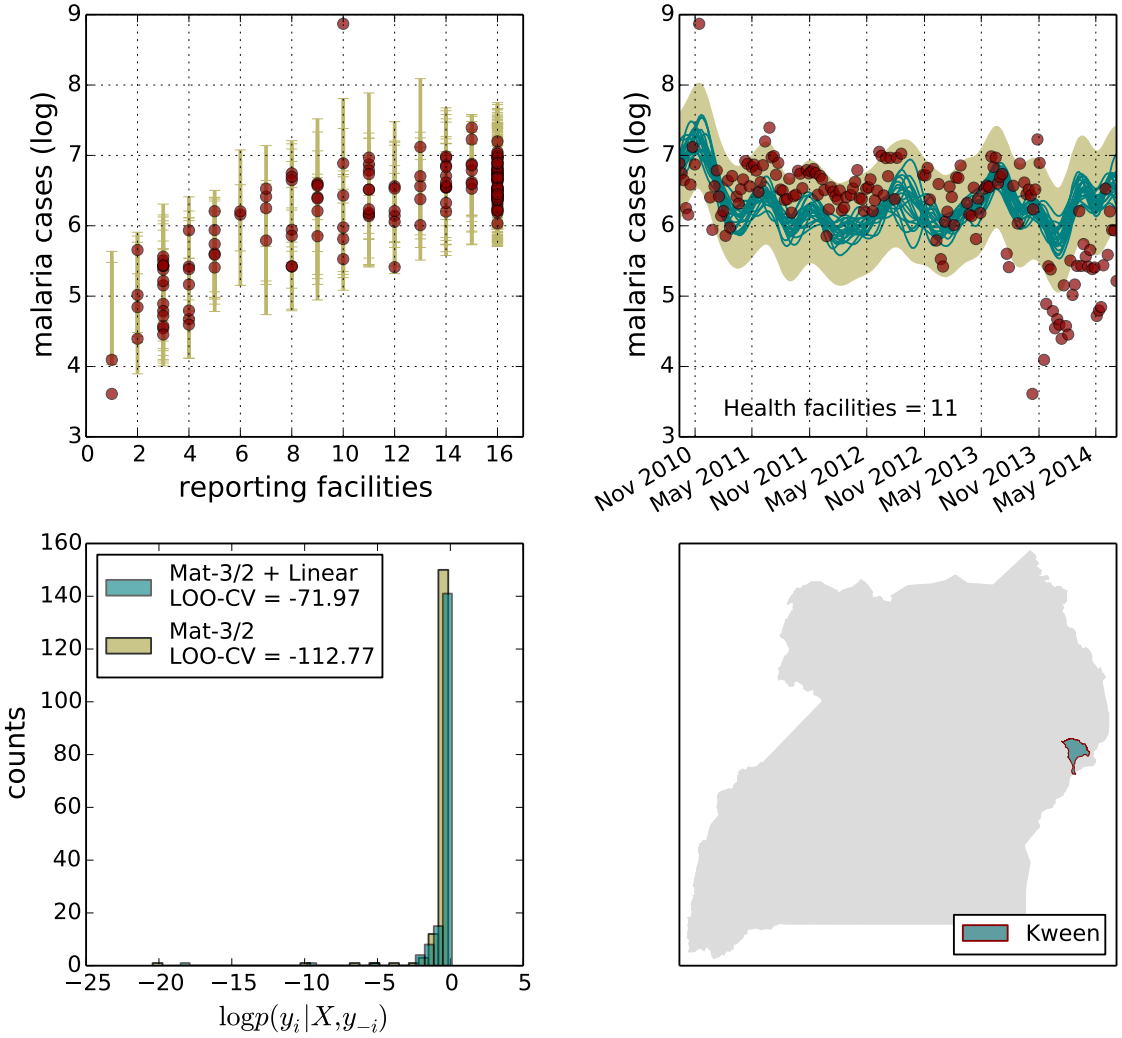


Fig. 5.5 Kernel selection for Kween district. Top left panel shows the number of malaria cases (log) against the number of reporting facilities. The vertical lines represent a 95% credibility interval for each data point. Top right panel shows the model fitted to the data (log) against time. Assuming a constant number of reporting facilities, the shaded area shows to the 95% credibility interval and the lines show random realizations of the latent function learnt. Bottom left panel shows the leave-one-out log-probabilities for each data point in each of the models compared.

$\dot{\boldsymbol{\Sigma}}$ and a Gaussian for $\dot{\boldsymbol{\mu}}|\dot{\boldsymbol{\Sigma}}$. The predictive distribution of $\mathbf{z}_i$ is then Student-t (Gelman et al., 2014). Figure 5.8 shows this approach applied to Kalungu and Gomba districts. The dashed lines in the middle plots show the original time series, and the red lines show the time series once the unlikely observations have been removed. The criteria

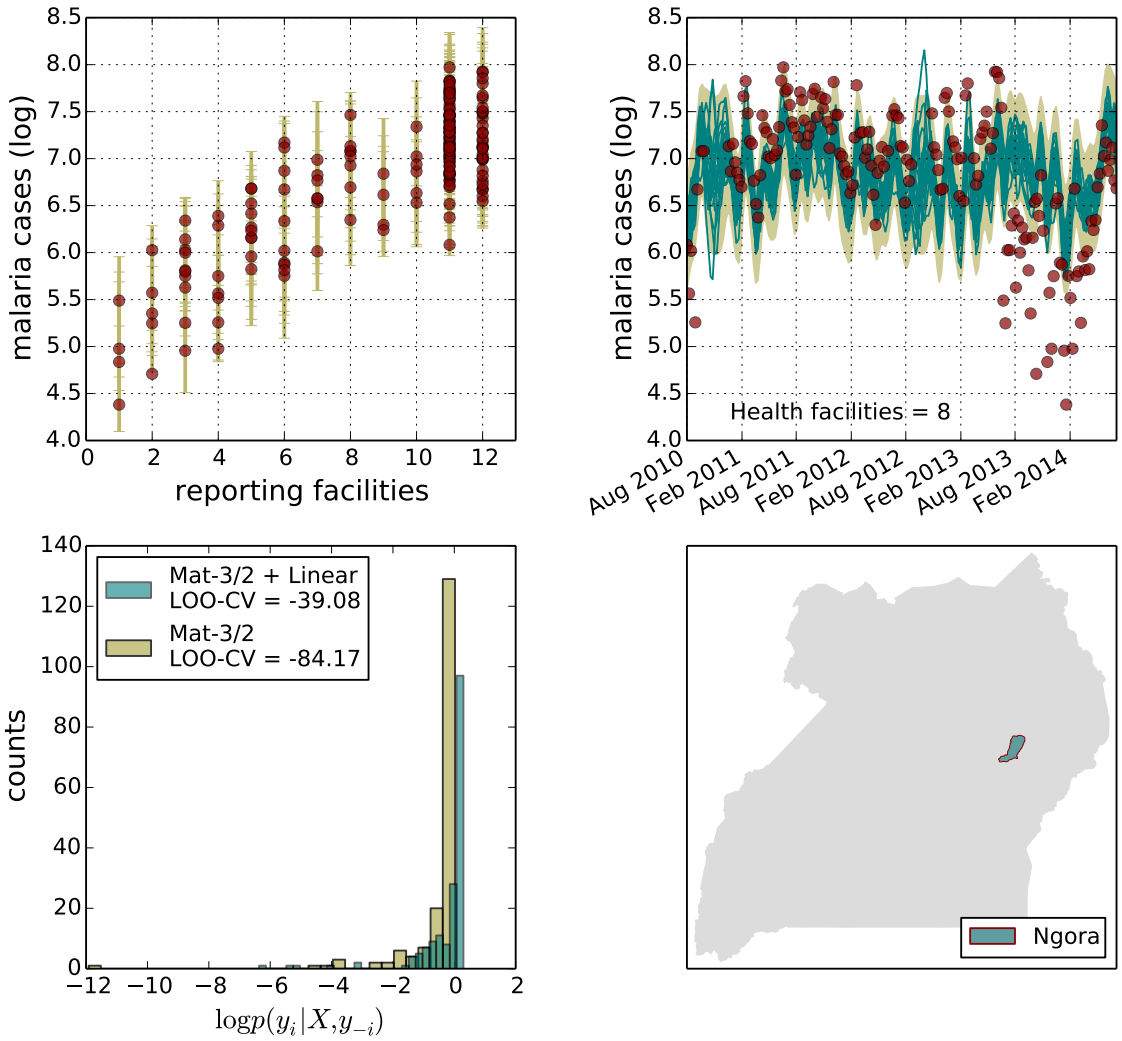


Fig. 5.6 Kernel selection for Ngora district. Top left panel shows the number of malaria cases (log) against the number of reporting facilities. The vertical lines represent a 95% credibility interval for each data point. Top right panel shows the model fitted to the data (log) against time. Assuming a constant number of reporting facilities, the shaded area shows to the 95% credibility interval and the lines show random realizations of the latent function learnt. Bottom left panel shows the leave-one-out log-probabilities for each data point in each of the models compared.

to define *unlikely* was to be outside the (rotated) ellipse $\mathcal{A}$ centered on $\dot{\boldsymbol{\mu}}$ and with semi-axis defined $3 \times \dot{\boldsymbol{\Sigma}}_{11}$ and $3 \times \dot{\boldsymbol{\Sigma}}_{22}$.

We are not interested in modelling data points affected by ζ_i, as it is just unstructured noise and it is not related with the actual disease. But at the same time, we are

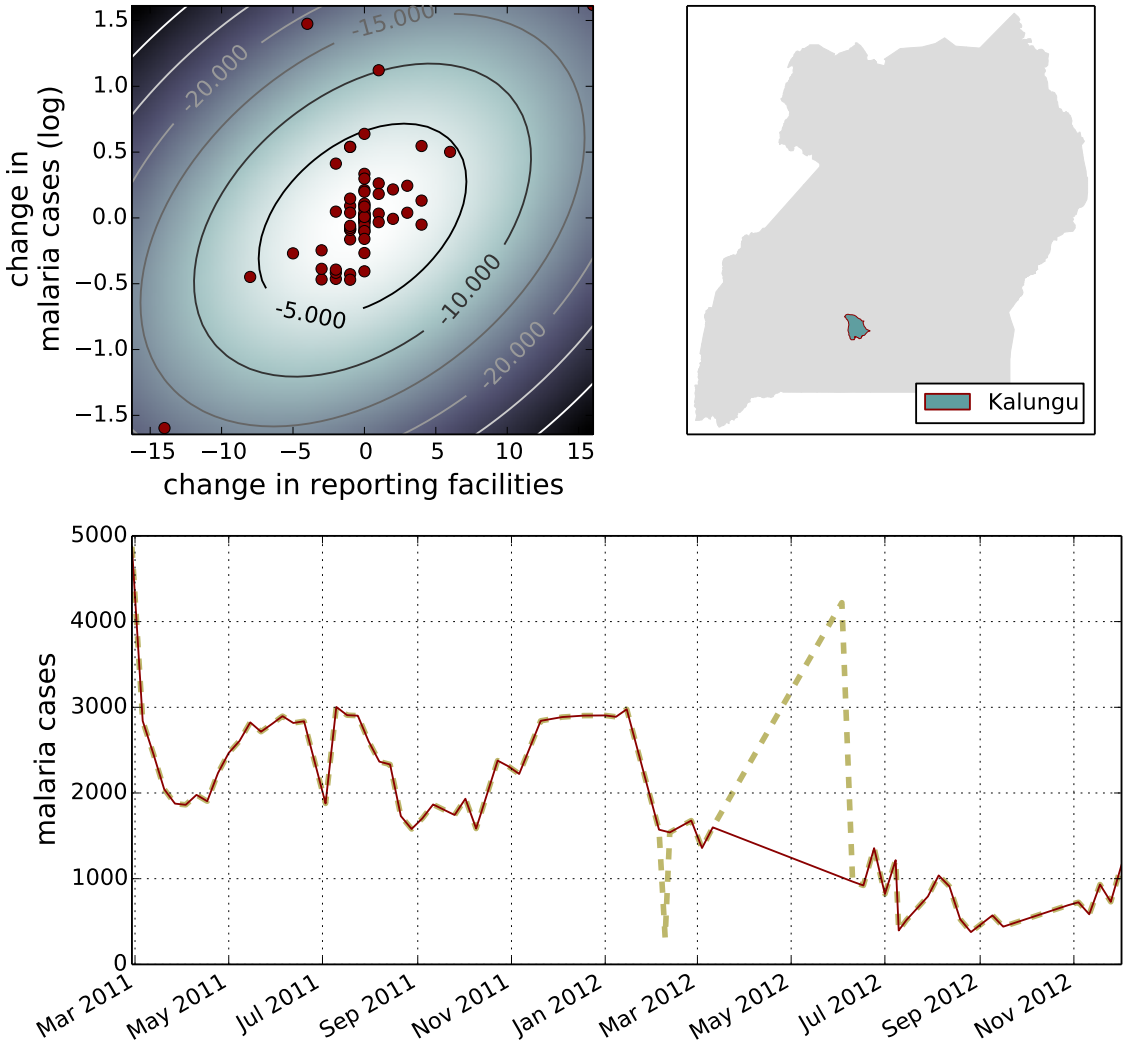


Fig. 5.7 Identification of potential errors in Kalungu. Top left panel shows the joint distribution of changes in malaria cases vs changes in reporting facilities. Bottom panel compares the HMIS data (dashed line) with the filtered data (solid line).

not willing to remove points without being confident of them being errors. It is a good practice not to do so, since if we remove a point that is not an error, we are limiting the learning of our model. A further inspection of these potential errors is needed to provide us a better understanding and confidence about the overall data consistency. Should the points outside the ellipse $\mathcal{A}$ be indeed atypical values in the time series, an

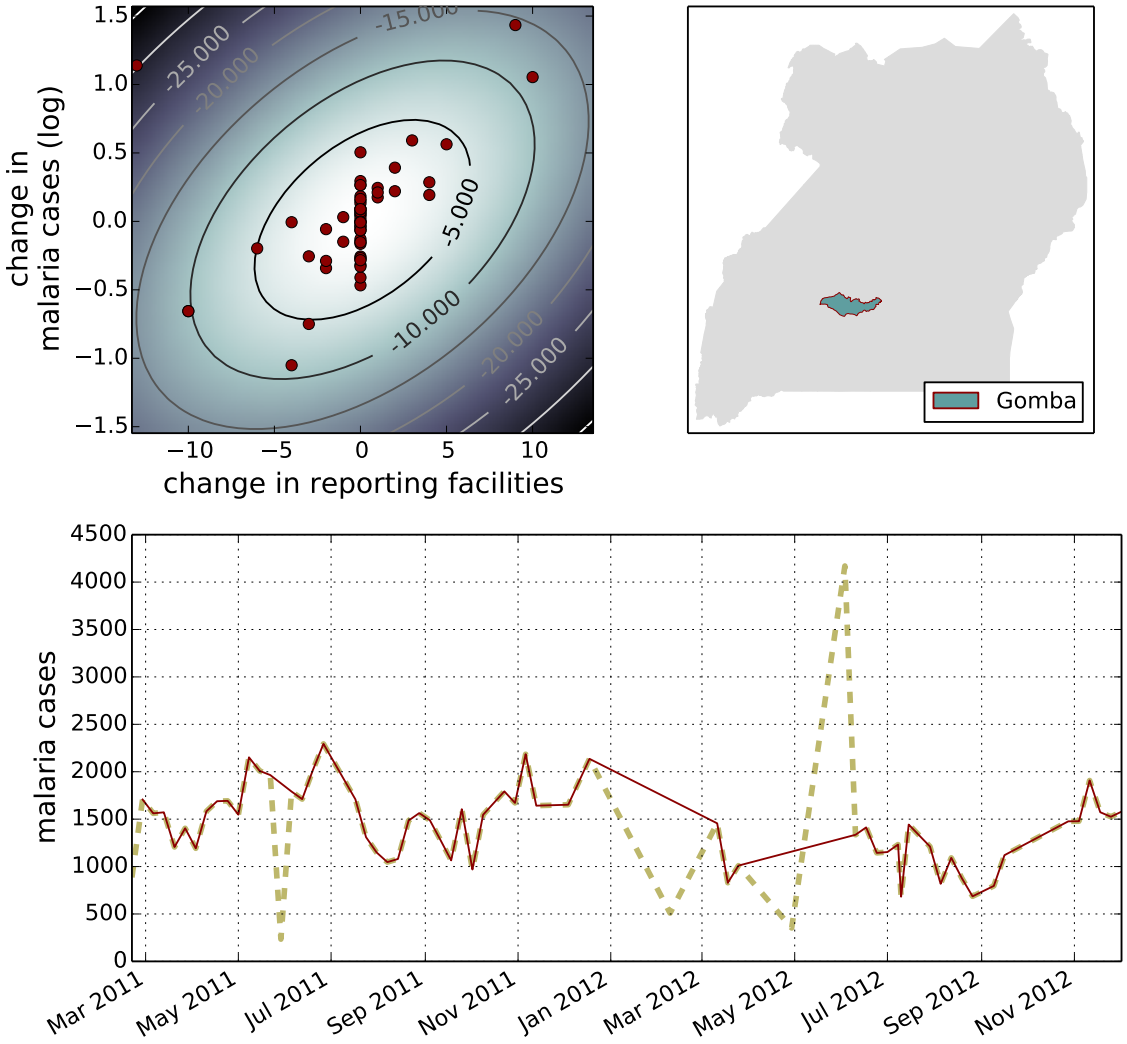


Fig. 5.8 Identification of potential errors in Gomba. Top left panel shows the joint distribution of changes in malaria cases vs changes in reporting facilities. Bottom panel compares the HMIS data (dashed line) with the filtered data (solid line).

homogeneous model like

$$\log y_i = f_{\mathbf{x}_i} + \epsilon_i \tag{5.8}$$

would be outperformed by a model like

$$\log y_i = f_{\mathbf{x}_i} + \epsilon_i \mathcal{I}_{\{\mathbf{z}_i \in \mathcal{A}\}} + \delta_i \mathcal{I}_{\{\mathbf{z}_i \notin \mathcal{A}\}}, \tag{5.9}$$

where $\delta_i \sim \mathcal{N}(0, \sigma^2_{\delta_i})$ has heterogeneous variance across observations, such that $\sigma^2_{\delta_i} > \sigma^2_\epsilon$; and $\mathcal{I}$ are index variables that make sure ϵ_i and δ_i are applied only when $\mathbf{z}_i \in \mathcal{A}$ and $\mathbf{z}_i \notin \mathcal{A}$, respectively. The equivalence of this model with (5.1) becomes clear if we note that $\delta_i = \zeta_i + \epsilon_i$. Some heteroscedastic models learn a functional form of the noise term across the input space $\mathbb{S}$ (Goldberg et al., 1998; Lázaro-Gredilla and Titsias, 2011; Tolvanen et al., 2014; Wu et al., 2014). However this is not the approach needed at this point. The noise term considered here is unstructured. Model (5.9) assumes homogeneity in all observations but a few, which have already been identified. The intention behind this approach is to have a flexible model allowed to weight differently some training instances. An example is shown in Figure 5.9. When the same noise variance is used across all observations, the GP fitted tries to yield trajectories that are equally close to each data point. When different noise variances are used, the model fits better those observations with the smaller noise. The trajectories yielded in the heteroscedastic model can go farther from (or even ignore) those observations with very large noise variance.

For each district, we compared models (5.8) and (5.9) using LOO-CV. Figures 5.10 and 5.11 show the models fitted for Nwoya and Bukwo districts. When adding an extra-parameter $\sigma^2_{\delta_i}$ for each potential outlier y_i, the overall variance σ^2_ϵ, used in most of the observations, has the opportunity to decrease in comparison to its value in the homogeneous model. The last means that the credibility interval for the non-outliers can shrink and thus reduce the uncertainty in the process. In general, adding an extra noise parameter for the atypical values increased the LOO-CV predictive probabilities. The results of every model fitted are presented in Appendix E.3.

5.6.5 Harmonization Across Different District Definitions

Some districts data have a structure that complicates comparability across time (see Figures 5.3a and 5.3b). On one hand the reporting facilities decrease each time the district is split. On the other hand, some reporting facilities are added to each new district. If we add up the information of the current districts and compare it to the parent district, we will see a dramatic increase in the infections. We hypothesize that the observed increment in the disease is artificial and it is due to an improvement in the health coverage when adding new facilities to the system[9]. The challenge now is to

[9]Proving an underestimation of malaria case-counts due to health coverage limitations in early periods goes beyond the goal of this work, as it would require information from other sources we do not have access to.

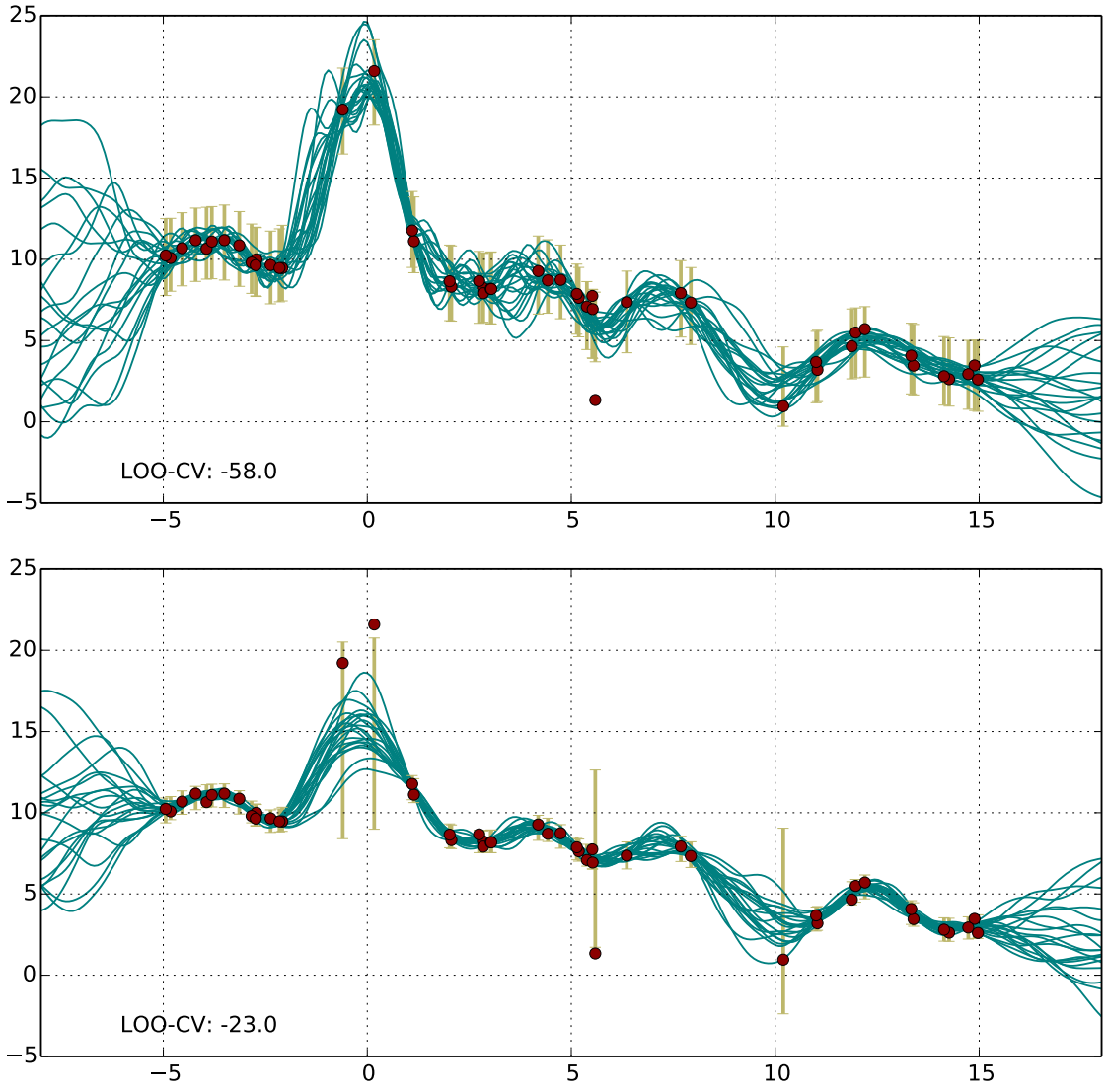


Fig. 5.9 Homoscedastic and heteroscedastic regression (toy example). Top panel shows the GP fit using the same noise variance for all observations. Bottom panel shows the GP fit using different noise variances for 4 observations. The dots indicate the observed points; the vertical lines show the predictive interval for the outputs; and the wiggling lines are random realizations of the latent function learnt.

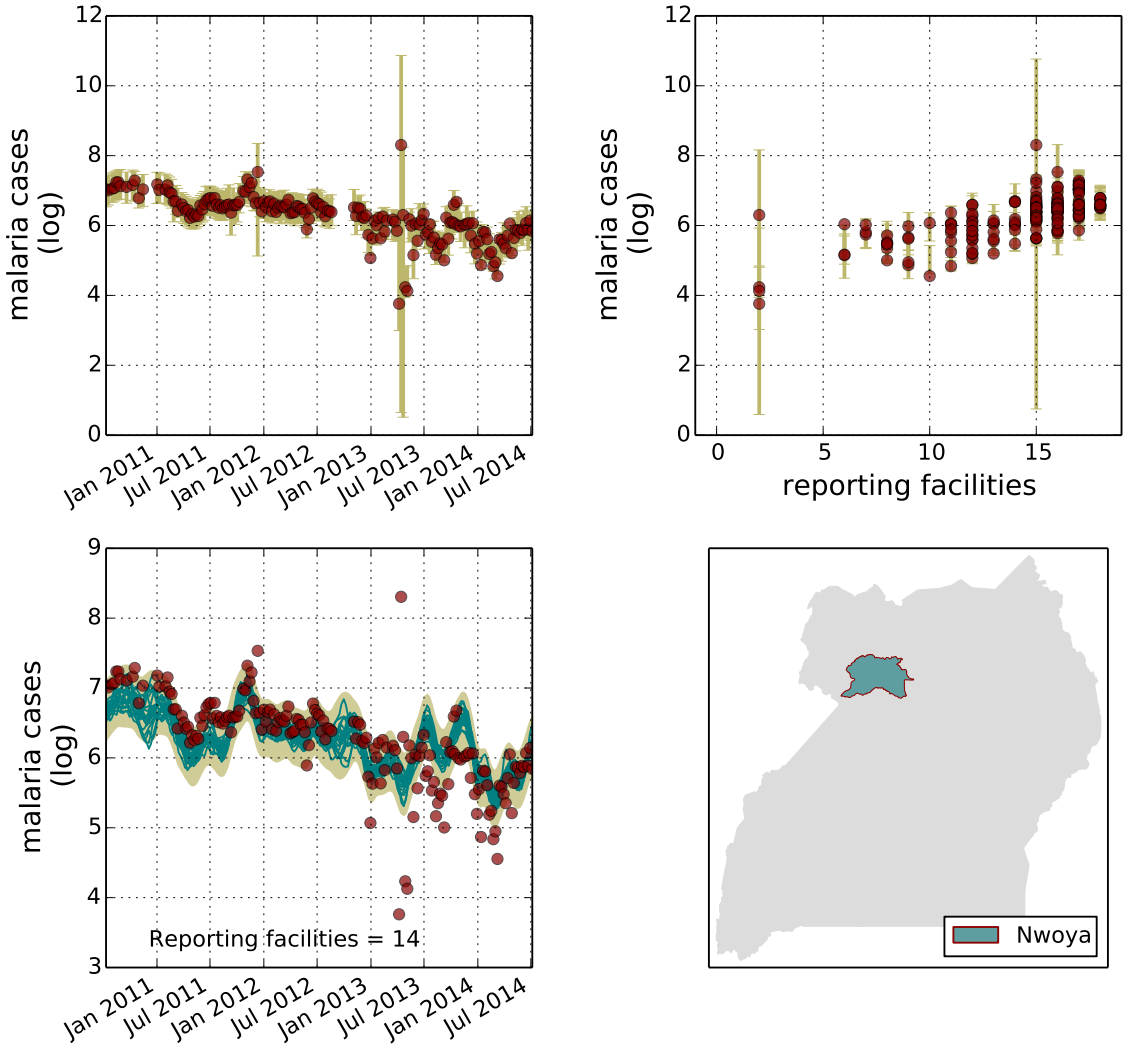


Fig. 5.10 Heteroscedastic model for Nwoya. Top panels show the number of malaria cases (log) vs time (left) and number of health facilities reporting (right). The vertical lines represent a 95% credibility interval predicted for each data point. Bottom left panel shows different simulations (lines) of the latent variable assuming a constant number of health facilities. The shaded are corresponds to the 95% credibility intervals of the reported data.

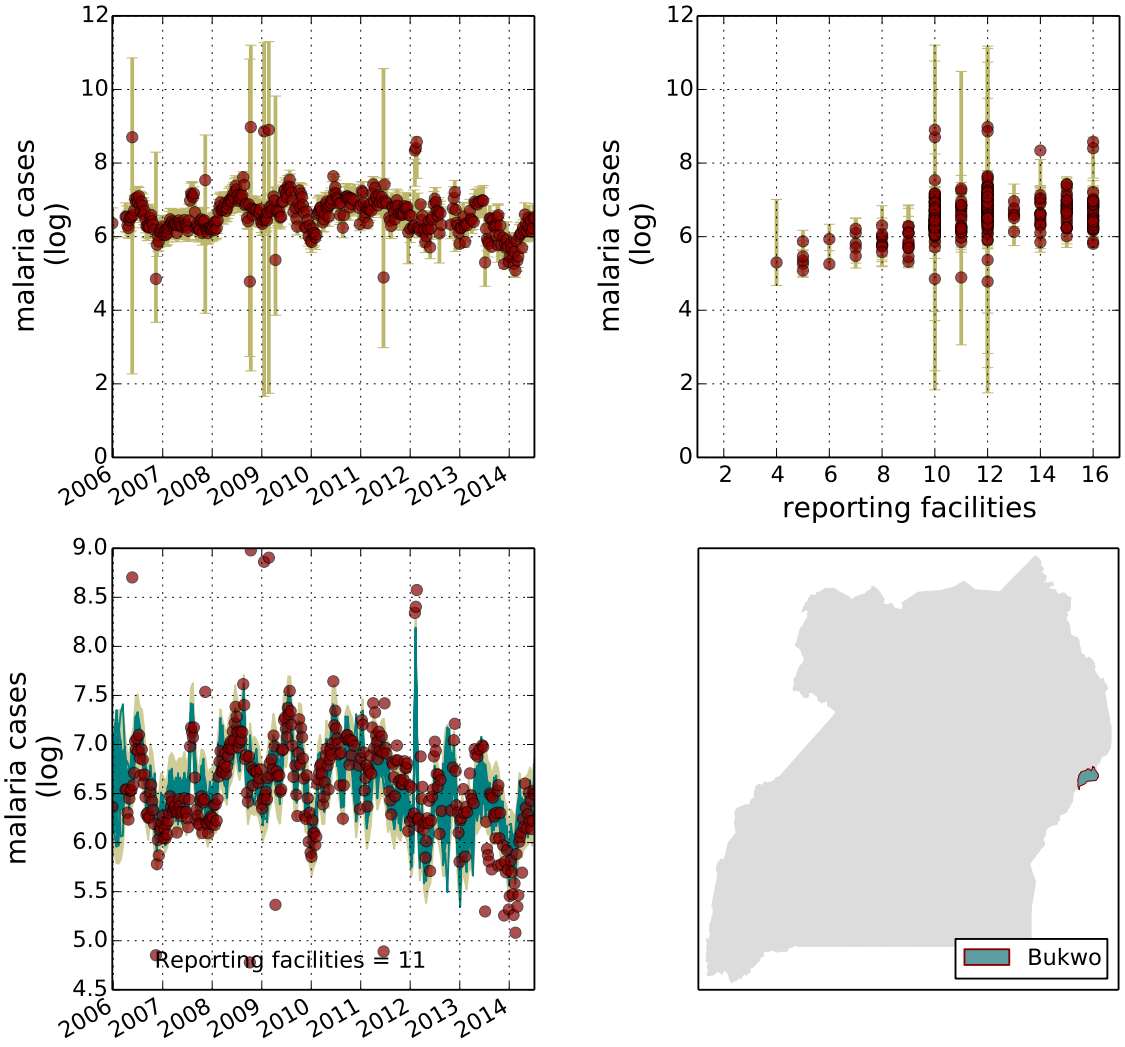


Fig. 5.11 Heteroscedastic model for Bukwo. Top panels show the number of malaria cases (log) vs time (left) and number of health facilities reporting (right). The vertical lines represent a 95% credibility interval predicted for each data point. Bottom left panel shows different simulations (lines) of the latent variable assuming a constant number of health facilities. The shaded are corresponds to the 95% credibility intervals of the reported data.

make data comparable across time. We propose a model, robust enough, to be used across the changes in the geoadministrative framework of the country.

A trivial model that explains the data of a district A just as a constant mean plus noise, i.e.,

$$y_{Ai} = \gamma_A + \epsilon_{Ai}, \tag{5.10}$$

can be defined as a GP with a kernel of the form

$$K(\mathbf{x}_i, \mathbf{x}_j) = \gamma_A^2. \tag{5.11}$$

Suppose that at some point A is split into A' and B, and that these two new districts are modelled in a similar way to (5.10). If there are no biases in the measurements of any of the districts, such that $\langle \epsilon_{\cdot i} \rangle = 0$ for all them, then it should be satisfied that

$$\gamma_A = \gamma_{A'} + \gamma_B. \tag{5.12}$$

This restriction in the mean parameters of the districts, can be worked out using a vector valued Gaussian process. Let $\mathbf{y}_i = (y_{Ai}, y_{A'i}, y_{Bi})^\top$, the corresponding kernel that satisfies the properties we ask for is defined as

$$\Gamma(\mathbf{x}_i, \mathbf{x}_j) = \left[\gamma_{\jmath(i)}\gamma_{\jmath(j)}\right], \tag{5.13}$$

where $\jmath(\cdot)$ is an index associated to any of the districts. Using the same principle, we can build a *nested-mean* kernel of a set of districts with a more involved tree structure, such as Masaka or Mbarara, which were split more than once.

In the context we are dealing with, we know there are biases in the measurements. To model this bias while keeping $\langle \epsilon_{\cdot i} \rangle = 0$, we add an additional term $\boldsymbol{\tau} = (\tau_A, \tau_{A'}, \tau_B)^\top$ to (5.10) as follows

$$\Gamma(\mathbf{x}_i, \mathbf{x}_j) = \left[\gamma_{\jmath(i)}\gamma_{\jmath(j)} + \tau_{\jmath(i)}^2 \mathcal{I}_{\{\jmath(i)=\jmath(j)\}}\right]. \tag{5.14}$$

The second term in the r.h.s. of (5.14) is added only when computed across data points of the same district. This is needed as the bias is related to the conditions specific to a particular district definition. For this exercise, we will assume that the main cause of bias is an incomplete health coverage in the earlier periods, so that the current observations have no bias, this is $\tau_A' = 0$ and $\tau_B = 0$.

Everything seems reasonable, except for the fact that so far we have been working on the log-space. An easy way of incorporating this vector-valued component with the

real-valued models constructed before is by defining

$$\log \mathbf{y}_i = \boldsymbol{\beta} + h_{\mathbf{x}_i} + \boldsymbol{\epsilon}_i, \tag{5.15}$$

where $\boldsymbol{\beta} = .5 \log(\boldsymbol{\gamma}^2 + \boldsymbol{\tau}^2)$, for $\boldsymbol{\gamma} = (\gamma_A, \gamma_{A'}, \gamma_B)$, and $h_{\mathbf{x}_i} = (f^A_{\mathbf{x}_i}, f^{A'}_{\mathbf{x}_i}, f^B_{\mathbf{x}_i})^\top$ is a vector-valued Gaussian process. The term $\boldsymbol{\epsilon}_i = (\epsilon_{Ai}, \epsilon_{A'i}, \epsilon_B)^\top$ in (5.15), is modelled in a similar way as we have proceeded so far, along this chapter.

The kernel of each component $f^*_{\mathbf{x}_i}$ is defined in the same way it was done for the single-district models (see Equation (5.8)). Let $K_A, K_{A'}$ and K_B be the corresponding kernels used for each district. All three kernels are easily embedded in (5.15) with a linear coregionalization kernel of the form

$$\Gamma(\mathbf{x}_i, \mathbf{x}_j) = K_A(\mathbf{x}_i, \mathbf{x}_j) \times \mathbf{e}_1\mathbf{e}_1^\top + K_{A'}(\mathbf{x}_i, \mathbf{x}_j) \times \mathbf{e}_2\mathbf{e}_2^\top + K_B(\mathbf{x}_i, \mathbf{x}_j) \times \mathbf{e}_3\mathbf{e}_3^\top, \tag{5.16}$$

where $\mathbf{e}_i$ is the i-th canonical basis vector of $\mathbb{R}^3$.

The more districts we handle within the same vector-valued process, the larger the computational burden. Here is where the use of sparse approximations becomes a great ally and helps overcoming the restrictions of large data sets. Nevertheless, up to this point we have been able to handle the computational costs using the full covariance and there has not been a need to use any approximation.

Figures 5.12 and 5.13, show two examples of how series are harmonized using the method proposed. In addition to the nested-mean kernel, the estimated data assumes a constant number of reporting facilities within each period. The estimated bias correction shown in the bottom left panel, correspond to the difference between the estimated mean of model (5.15) and the mean yield by the nested-mean kernel.

5.7 Addition of Environmental Variables

We have discussed how malaria has a strong link with the environment. If we could unravel this link and embed it into our model, we would be able to improve our characterization of the disease and even develop better tools for forecasting. In this last section of the chapter, we investigate if the addition of environmental data benefits our uncertainty measurement about HMIS data. We will use NDVI as a surrogate variable for the environmental conditions.

NDVI is computed from satellite images with a high spatial resolution. On the contrary, the spatial resolution of HMIS data is very poor, as it is aggregated at a district level. This restricts our analysis to be focused mainly on time, rather than

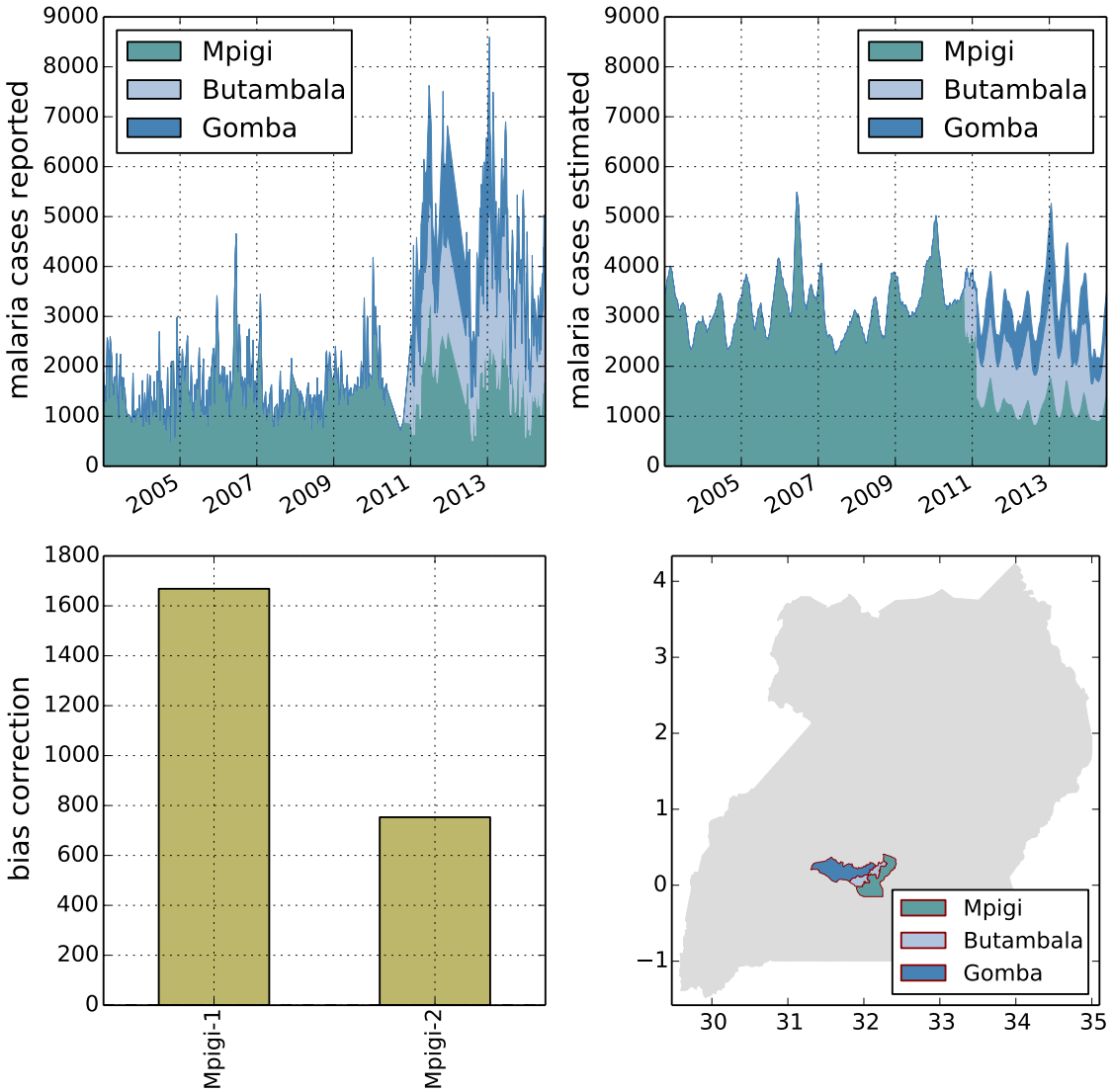


Fig. 5.12 Series harmonization of Mpigi, Butambala and Gomba. Upper left panel shows the observed data. Upper right panel shows the estimated mean after harmonizing the series. Bottom left panel shows the bias correction used on the previous definitions of Mpigi. Bottom right panel shows the districts location.

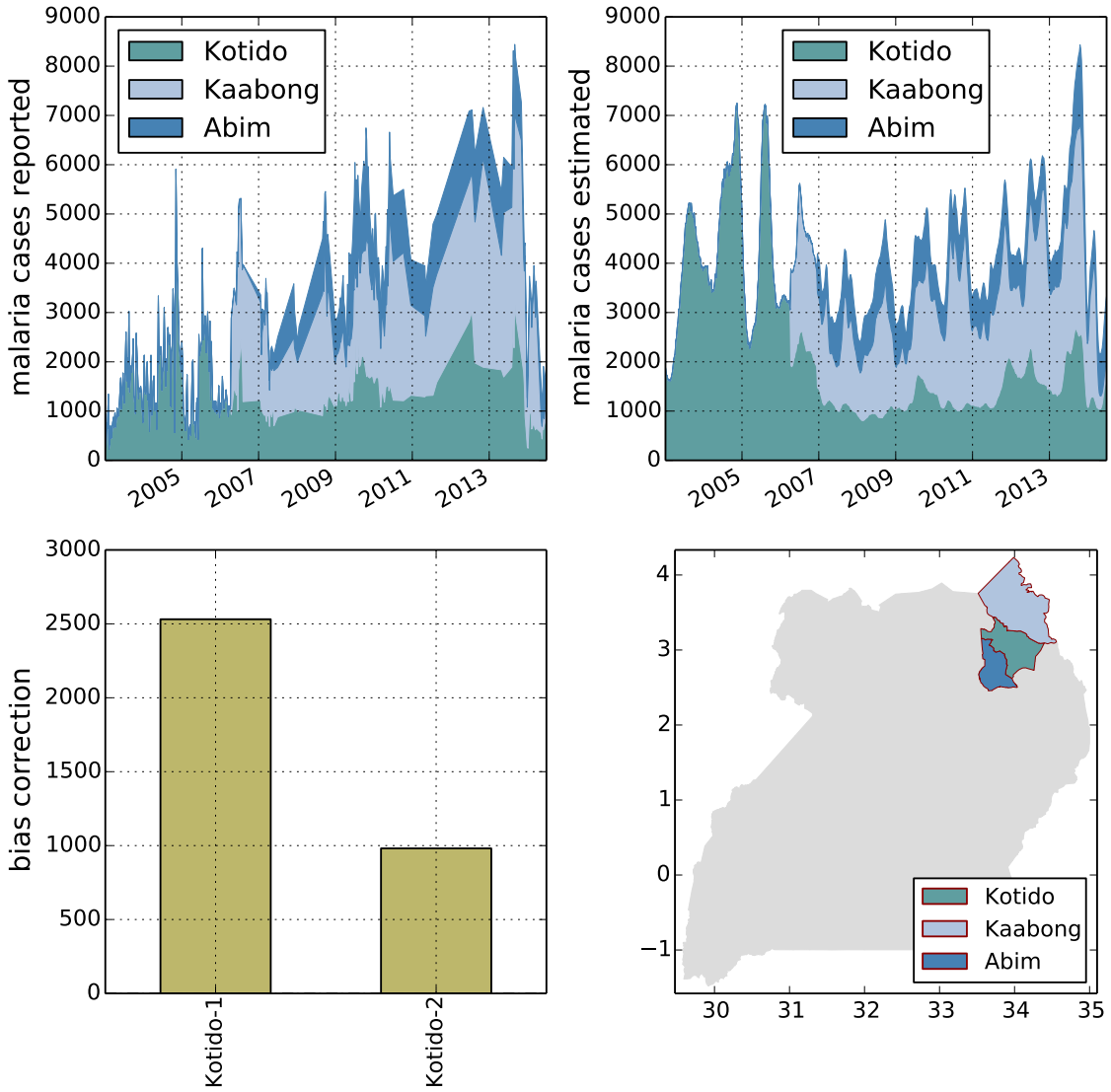


Fig. 5.13 Series harmonization of Kotido, Kaabong and Abim. Upper left panel shows the observed data. Upper right panel shows the estimated mean after harmonizing the series. Bottom left panel shows the bias correction used on the previous definitions of Kotido. Bottom right panel shows the districts location.

on time and space. From January 2008 to December 2012, we computed an average NDVI index for each district by weighting the pixel values across space according to the population estimates. Population estimates per pixel were obtained from the WorldPop project. A lag of two months on NDVI was used for associating it with HMIS (Ceccato et al., 2007; Haque et al., 2010). Only HMIS data produced under the current district framework was used.

There are different ways in which we can model the relation between HMIS and NDVI. We define and compare two approaches for doing it. As a benchmark or base model, let the disease case-counts y^H in a district be modelled as

$$\log y_i^H = f_{\mathbf{x}_i}^H + \epsilon_i^H, \tag{5.17}$$

where $f_{\mathbf{x}_i}^H$ is a GP. Similar to Section 5.6.4, we use a kernel function that takes time and the number of reporting facilities as inputs in the following way

$$K_{(0)}(\mathbf{x}_i, \mathbf{x}_j) = K_t(t_i, t_j) + K_r(r_i, r_j), \tag{5.18}$$

where K_t is an RBF kernel[10] and K_r is a linear kernel.

5.7.1 NDVI as Input

Equations (5.17) and (5.18) define an open loop system in which the number of malaria patients is explained by time and number of reporting facilities. The input dimension of such a model is easily expanded by adding a third kernel K_ω that takes NDVI values as inputs. Since HMIS and NDVI data are not observed at the same time points, we first need to estimate NDVI data at the HMIS time points and then use these estimates as inputs. Let's start by defining a GP regression model for NDVI as follows

$$-\log \frac{1 - y_i^E}{1 + y_i^E} = f_{\mathbf{x}_i}^E + \epsilon_i^E, \tag{5.19}$$

where y_i^E are NDVI data points and $f_{\mathbf{x}}^E$ has a kernel $K_e(t_i, t_j)$. The transformation in the l.h.s. of Equation (5.19) is used to expand the NDVI range values from $(-1, 1)$ to $(-\infty, \infty)$, so that a Gaussian likelihood can be used.

[10]The decision to use an RBF kernel instead of a Matérn, as we had been doing for most districts, is a matter of practicality. In the software used, the regression model with uncertain inputs, needed in this section, is only implemented when using RBF kernels.

Model (5.19) provides a set of interpolation points $\hat{\omega}_i$ that match the time points of HMIS data and can be used as inputs in the kernel function

$$K_{(1)}(\mathbf{x}_i^*, \mathbf{x}_j^*) = K_t(t_i, t_j) + K_r(r_i, r_j) + K_\omega(\hat{\omega}_i, \hat{\omega}_j), \tag{5.20}$$

for $\mathbf{x}_i^* = (t_i, r_i, \hat{\omega}_i)$.

It is important to notice that we are using NDVI estimates as inputs. We should be careful not to dismiss the uncertainty of these estimates, and propagate it across the estimates of malaria. As discussed in the previous chapter, input uncertainty is easily handled within the GP variational framework, where we use a distribution $q(\mathbf{x}_i^*)$ over the inputs.

5.7.2 HMIS and NDVI in a Vector-Valued GP

The middle step in which NDVI points are interpolated, to be later used as inputs, can be avoided if, instead of increasing the input dimensionality of y^H, we use a multi-task setting that relates y^H and y^E. In this system, the relation between the outputs y^H and y^E is analyzed directly through a vector-valued GP with the following kernel function

$$\Gamma_{(2)}(\mathbf{x}_i, \mathbf{x}_j) = K_t(t_i, t_j) \times \mathbf{e}_1\mathbf{e}_1^\top + K_r(r_i, r_j) \times \mathbf{e}_2\mathbf{e}_2^\top + K_{t'}(t_i, t_j) \times \mathbf{B}, \tag{5.21}$$

where $\mathbf{e}_i$ is the i-th canonical basis vector of $\mathbb{R}^2$ and $\mathbf{B} \in \mathbb{R}^{2\times 2}$ is a positive definite matrix. Notice that kernels K_t and K_r are private to y^H, while $K_{t'}$ is shared across both outputs and weighted by the coregionalization matrix. To ensure $\mathbf{B}$ is positive definite, and thus a valid kernel, we define it as

$$\mathbf{B} = \mathbf{w}\mathbf{w}^\top + \boldsymbol{\kappa}\mathbf{I}, \tag{5.22}$$

where $\mathbf{w}$ and $\boldsymbol{\kappa}$ are vectors in $\mathbb{R}^2$ and $\mathbf{I}$ is the identity matrix in $\mathbb{R}^{2\times 2}$.

5.7.3 Approaches Comparison

By no means the models described by the kernels (5.20) and (5.21) are equivalent. One represents an open loop system and the other a closed loop system. In the latter, both HMIS and NDVI are explained using the available information from the other variable. This means that NDVI estimates will depend on HMIS estimates. In the open loop system, it is only HMIS estimates that can be affected by NDVI.

For every district, we compared the performance of these models using 5-fold cross-validation[11]. The kernel in Equation (5.20) was used in a sparse variational regression, where the number of inducing inputs was one third of the total of observations[12]. The open loop system's CV scores tend to be worse than the benchmark ones. The closed loop systems performance is in general similar to the benchmark, with just a few cases being better. Figure 5.14 shows a comparison based on the CV scores.

We searched for a pattern in those cases where the closed loop system model outperforms the benchmark. We looked for features that could tell why NDVI benefits HMIS prediction in some districts and not in others. No conclusive results were found. For instance, we found no relation between the districts location or elevation with the model's performance when including NDVI[13] (see Figures 5.15 and 5.16). Disappointingly, the explanation is somewhere else. If the differences in the CV scores with respect to the benchmark is indeed related to the contribution of NDVI data, we should expect to see a relation between these differences and the values of the coregionalization matrix. In particular with the statistic

$$\rho = \frac{\mathbf{B}_{01}}{\sqrt{\mathbf{B}_{00}\mathbf{B}_{11}}}. \tag{5.23}$$

Figure 5.17 shows a plot of the CV differences (closed loop - benchmark) vs ρ. High difference values do not match high ρ values. In fact, the relation is the opposite. Values of ρ close to one are associated to differences close to zero and the largest differences are associated to values of ρ close to zero. In coregionalized regression problems, $\rho \approx 0$ means independence across outputs. The conclusion is that the improvement in the model fit does not come from adding NDVI information, but simply from using and additional kernel $K_{t'}$. Figure 5.18 shows a comparison of HMIS and NDVI series, for those districts with the largest CV differences. In all of these cases, there is no correlation between HMIS and NDVI.

We modified the base model by adding a second kernel for time $K_{t'}(t_i, t_j)$ and compared it to the base and closed loop models. As it is shown in Figure 5.19, the results between the closed loop model and the modified base model have very similar

[11] We do not use LOO-CV due to the difficulty of computing the LOO predictive probabilities when using sparse approximations. See Appendix D.

[12] Since the data series are not that large, there is no need to use a very sparse model. When the number of observations was less than 60 points, we decided to use a number of inducing inputs that matched the number of observations.

[13] There have been previous efforts to understand how the relation between malaria cases and climate is affected by spatial differences (Manh et al., 2011; Stern et al., 2011).

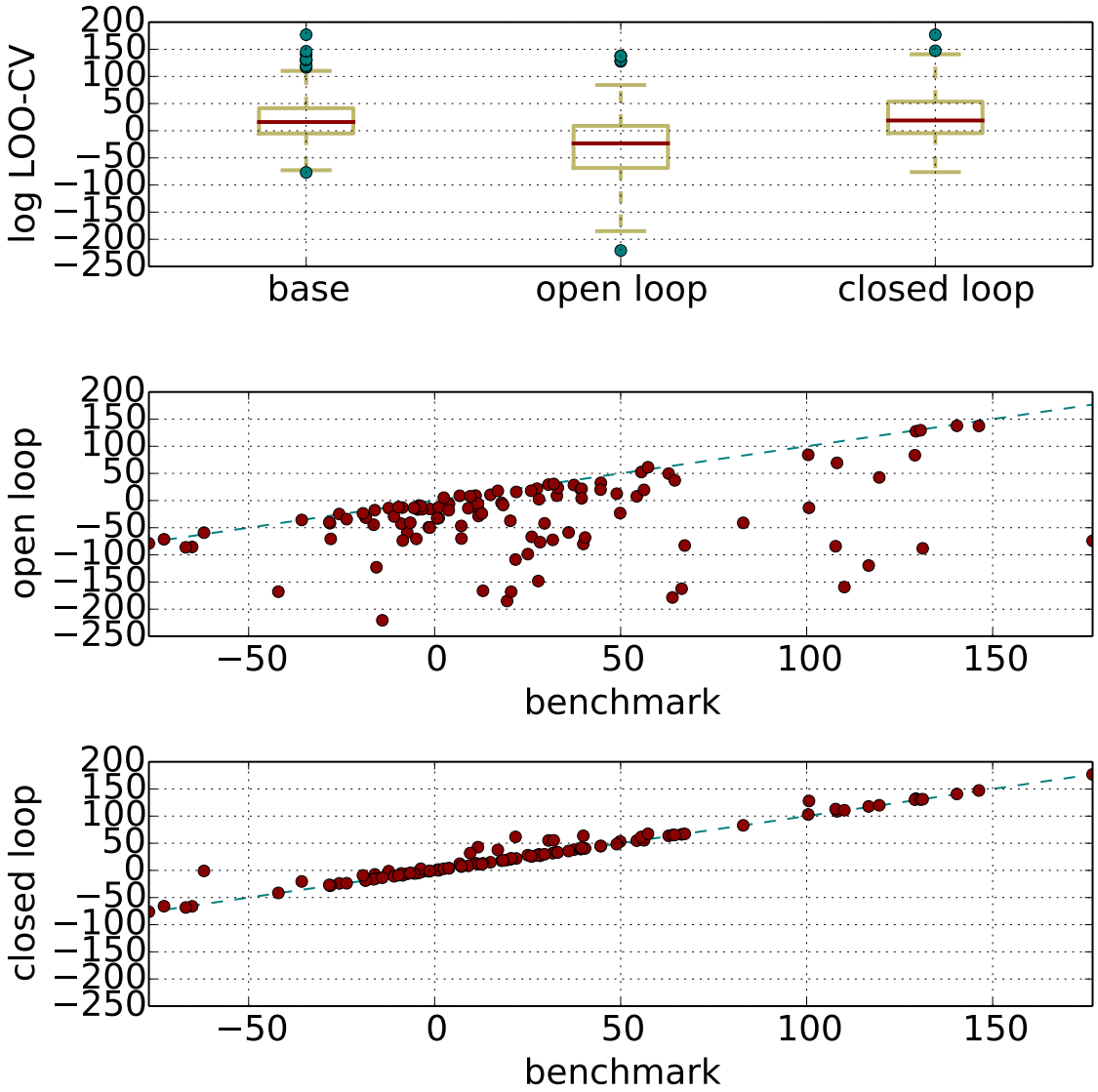


Fig. 5.14 NDVI in a closed and open loop systems. Top image compares CV results across the different approaches defined. Middle image compares CV scores per district in the open loop system vs the benchmark model. Bottom image compares CV scores per district in the closed loop system vs the benchmark model.

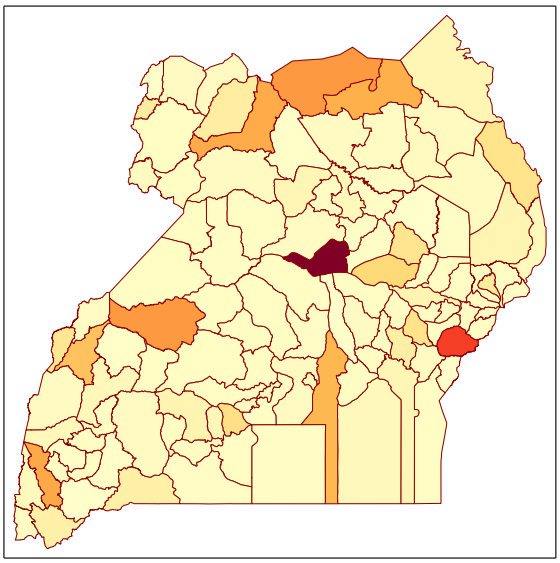


Fig. 5.15 CV differences per district. Darkest colors correspond to higher differences between closed loop system and the benchmark.

results. The bottom line of this similarity is that NDVI data is not helping reduce our uncertainty about HMIS data.

5.8 Final Comments

In this chapter, we have faced some of the challenges of assimilating real data into the model. Our approach was based on designing noise models and kernel functions that represent an *idealized* generating process of the HMIS records. We also showed how outliers can be assessed by contrasting an homoscedastic and heteroscedastic noise models. For this task, the heteroscedastic model does not need to represent a functional form of the variance. The underlying assumption is that large variance occurrences are independent from each other. Variations with a functional structure were modelled by combining different kernels.

The use of a nested-mean kernel provides promising results for harmonizing a series under a changing framework, like the administrative borders of the districts. Nevertheless, with the information available we have no means to verify if the backward estimation is actually correct. It is still pending to estimate the bias associated to the latest district framework. Also, we have not worked on the harmonization of the number of reporting health facilities. At the moment we have just shown results using the average number of facilities per district. We explored the benefit of incorporating

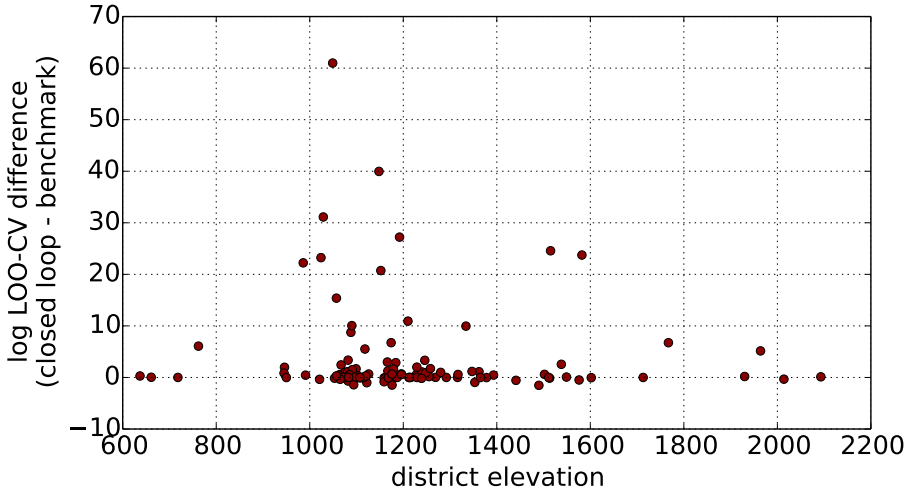


Fig. 5.16 CV differences vs district elevation.

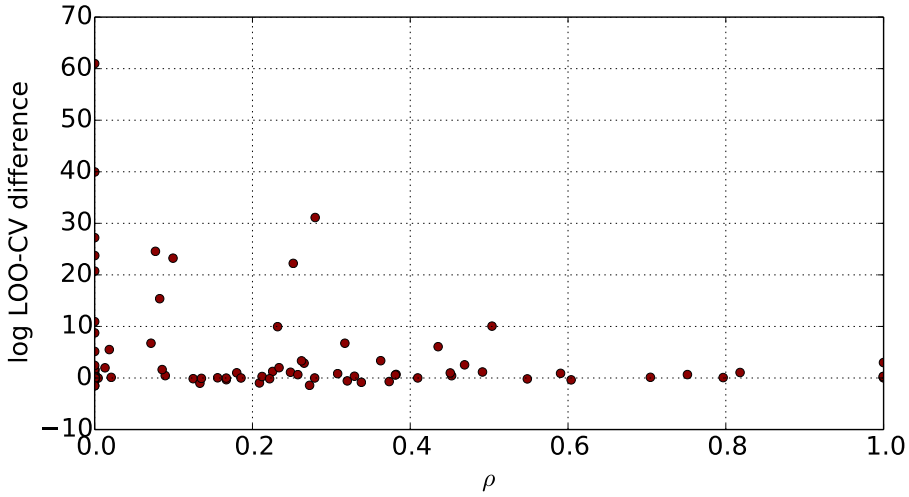


Fig. 5.17 CV differences vs ρ values. Highest ρ values occur when CV is close to zero.

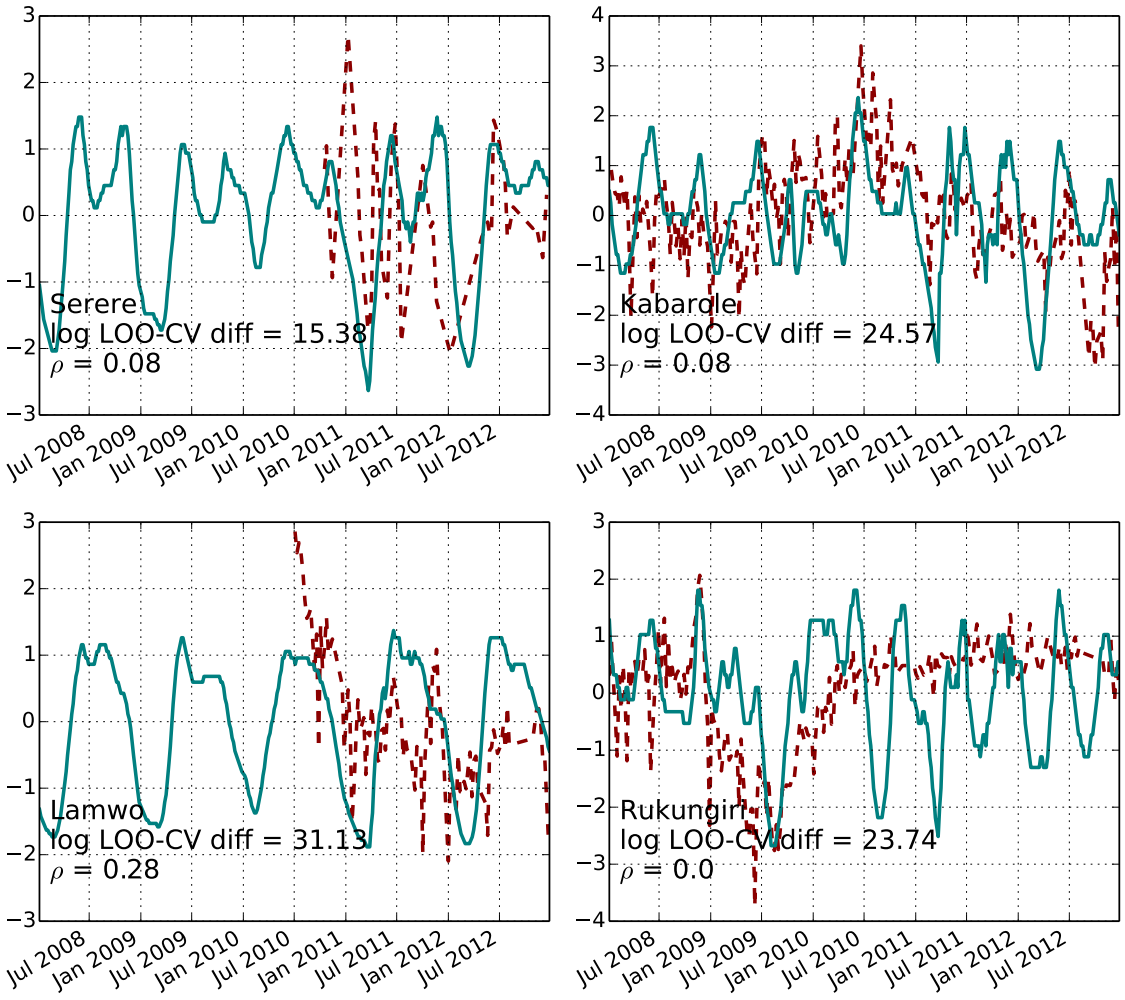


Fig. 5.18 HMIS and NDVI data (standardized). The dashed lines corresponds to HMIS data and solid lines correspond to NDVI data. The districts shown are the ones with highest CV difference with respect to the benchmark.

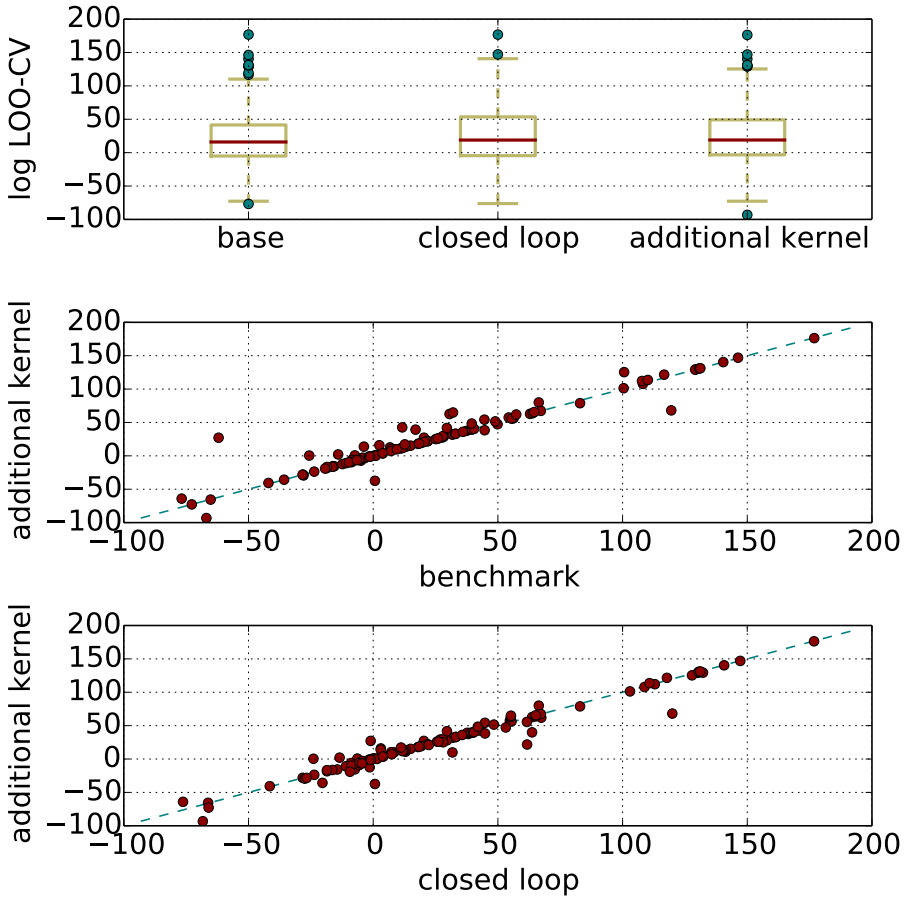


Fig. 5.19 Addition of $K_{t'}$ to the base model. Top image compares CV results across the base model, the close loop system model and the base model modified by adding an extra kernel. Middle image compares CV scores per district in the modified base model vs the benchmark model. Bottom image compares CV scores per district in the modified base model vs the closed loop system.

environmental data to the model, to improve the estimation of malaria case-counts. We could not find a clear pattern of association between NDVI and HMIS data. The only improvement in the fit came from using an additional kernel in the regression model.

In the next chapter, we will analyze more closely a covariance structure that combines two kernels on the time input to model malaria case-counts. We will use such kernel construction to monitor the disease dynamics in real time.

Chapter 6

Monitoring System of Malaria Case-Counts

Interventions to prevent and treat malaria will be successful depending on how well the disease can be anticipated and how fast the population reacts to it. We have worked on removing noise and biases from the data observed. The motivation is to have more accurate data, so that analysts and decision makers have a better starting point. In this section we go forward and build a tool to aid real-time analysis of health population across the country. The interest at this point is not forecast, but to characterize the population health at a specific time.

We have to agree that although the work in the previous section could make data comparable across the years and district definitions, it is still far from being an accurate measurement of the disease in the country. We have no means of quantifying estimation errors of malaria cases due to a lack of patient testing. We are also aware that health coverage and the varying number of reporting units would require field surveys or a closer work with the health authorities to be fully understood. Then the question is how to use what we have gained so far? How to contribute given the current circumstances?

Time series analysts frequently break down a series into different components, like trend and seasonal effects (Baxter and King, 1999; Cleveland and Tiao, 1976; Hyvärinen and Oja, 2000). In this sense, Gaussian process (GP) models are a natural approach for analyzing functions that represent time series. By combining different covariance kernels (via additions, multiplications or convolutions) into a single one, a GP is able to describe more complex functions. Each of the individual kernels contributes by encoding a specific set of properties or pattern of the resulting function (Durrande et al., 2013).

We propose a monitoring system for communicable diseases based on Gaussian processes. This methodology is able to isolate the relevant components of the time series and study the short term variations of the disease[1]. The output of this system is a graphical tool that discretizes the disease progress into four phases of simple interpretation.

6.1 Method Used

Lets say we have data generated from the combination of two independent source signals as the ones shown in Figure 6.1a. Usually we are not able to collect data coming from the source signals separately. In fact, we do not even observe the combined signal, but a corrupted version of it (see Figure 6.1b). Suppose that the smooth signal represents a long term trend component and that the sinusoidal signal represents a seasonal component. For an observer, the oscillations of the seasonal component masks the behaviour of the long term trend. This might pose a difficulty if the observer wants to know whether the trend is increasing or decreasing. In a similar way, the observer might be interested in the seasonal component isolated from the trend. The last, is a common practice in economics and finance, where business recession and expansion periods are determined by analyzing the cyclic component of a set of indicators (van Ruth et al., 2005). This way, the cyclic component tells if an indicator is above or below its trend, and its simple differences tell if it is increasing or decreasing. We propose a similar approach to monitor time series of disease case-counts, but in our case, we will use a non-parametric approach.

To extract the original signals, the observed data can be modelled using a GP with a combination of kernels, say exponentiated quadratics, one having a shorter lengthscale than the other. Figures 6.1c and 6.1d shows a model of the combined and independent signals. We also use a vector-valued GP to model directly the derivative of the time series, rather than using simple differences of the observed trend. As a result, we are able to provide uncertainty estimates about the speed of the changes around the trend. Our approach is based on modelling linear functionals of an underlying GP Särkkä (2011). If $h_{\mathbf{x}} = (f_{\mathbf{x}}, \partial f_{\mathbf{x}}/\partial x_i)^\top$, its corresponding kernel is defined as

$$\Gamma(\mathbf{x}_i, \mathbf{x}_j) = \begin{bmatrix} K(\mathbf{x}_i, \mathbf{x}_j) & \frac{\partial}{\partial x_j} K(\mathbf{x}_i, \mathbf{x}_j) \\ \frac{\partial}{\partial x_i} K(\mathbf{x}_i, \mathbf{x}_j) & \frac{\partial^2}{\partial x_i x_j} K(\mathbf{x}_i, \mathbf{x}_j) \end{bmatrix}. \tag{6.1}$$

[1]The methodology could be used the other way around, to study long term variations after removing the short term component.

If K is an exponentiated quadratic kernel, the block components of the vector-valued kernel defined in (6.1) are expressed as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\ell^2}\right), \tag{6.2}$$

$$\frac{\partial}{\partial x_i} K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\ell^2}\right)\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|}{\ell^2}\right), \tag{6.3}$$

$$\frac{\partial}{\partial x_i x_j} K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\ell^2}\right)\left(\frac{1}{\ell^2} - \frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{\ell^4}\right). \tag{6.4}$$

In most multi-output problems, observations of the different outputs are needed to learn their relation. Here, the relation between $f_{\mathbf{x}}$ and its derivative is known beforehand through the derivative of K. Thus $\partial f_{\mathbf{x}}/\partial x_i$ can be learnt by relying entirely on $f_{\mathbf{x}}$. For the signals described above, Figures 6.1e and 6.1f show the corresponding derivatives computed using a kernel like (6.1). The derivatives of the long term trend are computed with high confidence, while the derivatives of the seasonal component have more uncertainty. The last is due to the magnitude of the seasonal component relative to the noise magnitude.

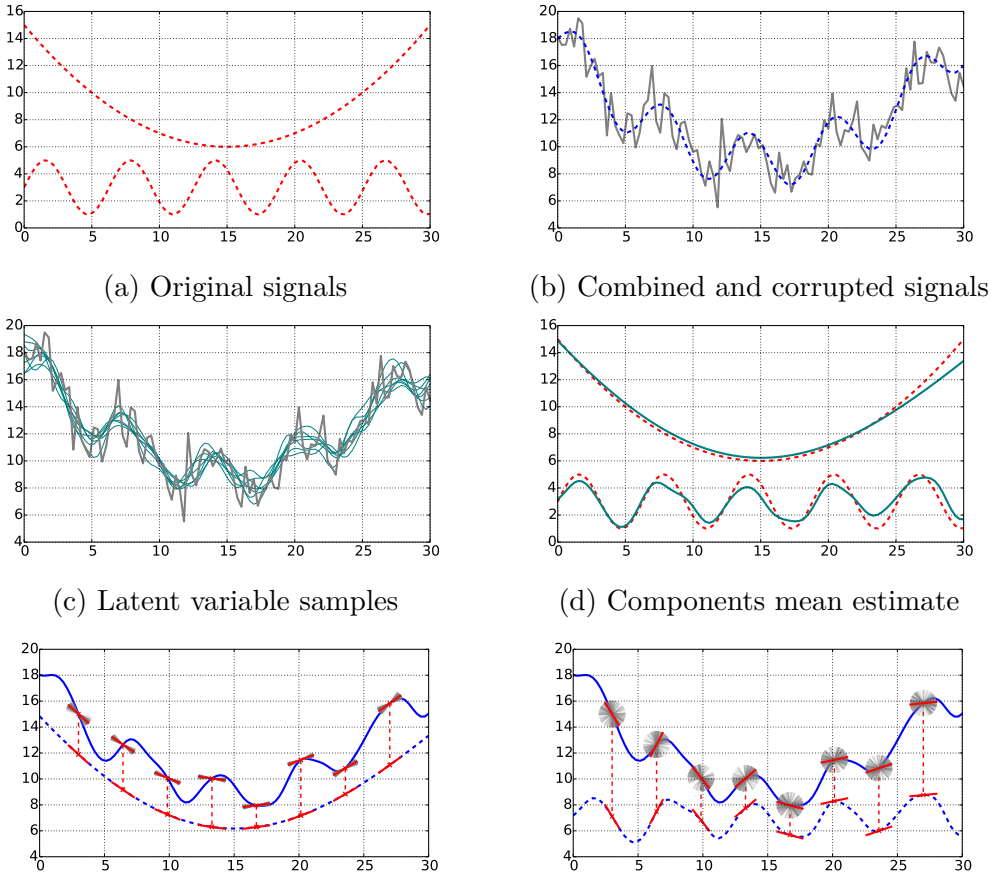
6.2 Uganda Case

For this exposition we focus on Kabarole district, but provide snapshots of the monitoring system for all the country. As we saw in Section 5, the number of reporting hospitals is not consistent across time. This variation is prone to create artificial trends in the observed data. Hence, the underreporting effect has to be estimated to be removed.

For this application, we used a linear kernel to model the effect of the number of reporting hospitals and a combination of exponentiated quadratic kernels to explain long and short term variations of the disease across time[2].

Figure 6.2a shows the trend and short term component of the number of malaria cases. Variations of a case-counts around the trend represent short term changes in the population health. Outbreak detection and control of non-endemic diseases take place in this time frame. For endemic diseases, this variation can be associated to seasonal factors (Hay et al., 1998). Quick response actions, such as distribution of medicine and allocation of patients to health centers, have to take place in this time regime to be

[2]If a seasonal effect in the series was clear, we could use a periodic kernel to model better the short term variations.



(a) Original signals

(b) Combined and corrupted signals

(c) Latent variable samples

(d) Components mean estimate

(e) Long term trend derivatives

(f) Seasonal component derivatives

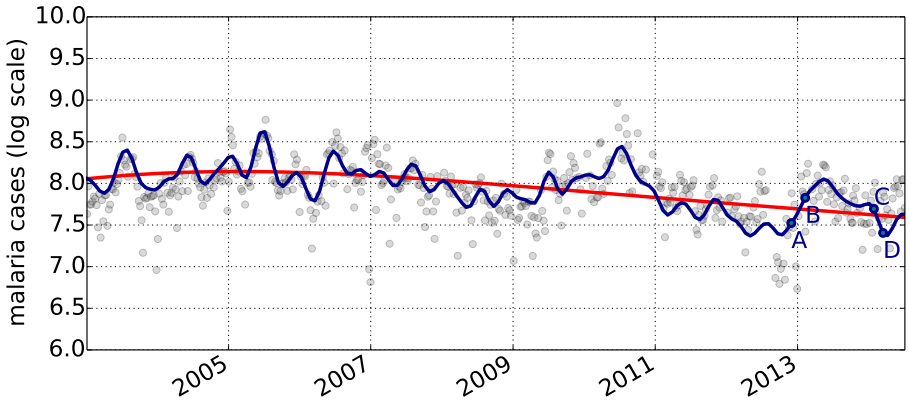
Fig. 6.1 Series decomposition. Panel (a) shows two independent signals. Panel (b) shows the combination of both signals (dashed line) and a distorted signal after adding some noise (solid line). Panel (c) shows latent variable samples representing the combined signal (thin lines). Panel (d) compares the mean estimate of each component (solid line) with the original signals (dashed line). Panels (e) and (f) show the components derivatives. Tangent lines to the individual components are shown in red. The solid blue lines represent the mean estimate of the composed signal. The gray lines are random realizations of process derivative. For comparison, the estimates of the individual signals (dashed lines) are shown below the composed signal.

effective. The short term variations can be classified in four phases as shown in Figure 6.2b (values are standardized). The upper left quadrant represents case-counts below the trend, but increasing; the upper right quadrant represents an case-counts above the trend and expanding; the bottom right quadrant represents case-counts above the trend, but decreasing; and the bottom left quadrant represents case-counts below the trend and decreasing.

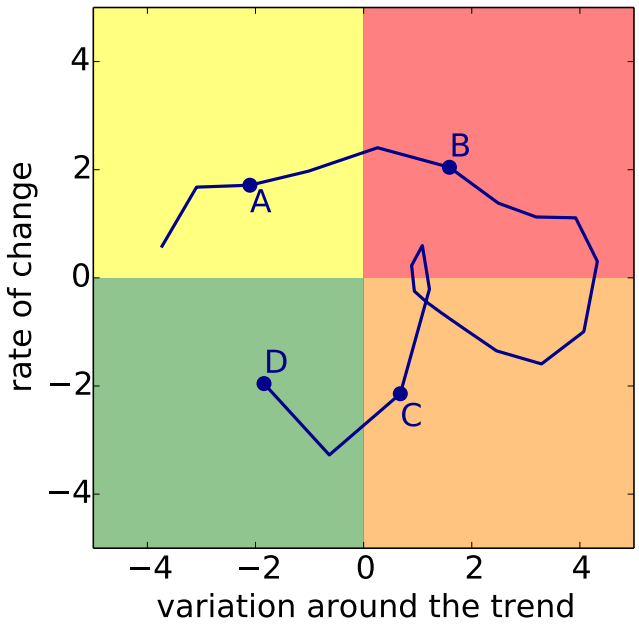
This tracking system of short term variations is independent of the order of the disease counts, and can be used to monitor the infection progress in different districts. It is easy to identify districts where the disease is being controlled or where the infection is progressing at an unusual rate. Figure 6.3 shows the monitoring system on the whole country during 4 consecutive weeks. Those districts where the variation coefficient of both the process and its derivative are less than 1 (meaning a weak signal vs noise) were left in gray color.

6.3 Final Comments

We have proposed a disease monitor based on vector-valued Gaussian processes. Our approach is able to account for uncertainty in both the level of each component and the direction of change. The simplicity for doing inference with this model is not compromised by the use of a vector-valued approach. The model can be benefited if spatial information is available and encoded in the kernel function. Further research is needed to explore the benefits of this model in practice. We expect that an analysis from this perspective can add situational awareness and contribute to interventions planning and resources allocation when facing infectious diseases.



(a) Trend and short term variations



(b) Cycle phases

Fig. 6.2 Malaria case-counts tracker in Uganda. Panel (a) shows the long term trend (red line) and the short term variations (blue line). Gray bullets represent the observed records. Panel (b) shows a tracking system of the short term variations. The bullets A-D correspond to the same time points in both panels.

Fig. 6.3 Disease monitor for all districts. Gray color means that the variation coefficient of both the short term process and its derivative are less than 1.

Chapter 7

Conclusions

Many things were learnt along the pages of this thesis and the hours or work they summarize. Here are listed the ones that seem to be the most relevant.

1. Covariance kernels are a well known structure able to encode non-linear patterns into a probabilistic model. By combining different kernels, signal decomposition can be easily handled.

2. We developed a sparse variant of the expectation propagation algorithm. This allows modelling data with different noise assumptions with a variational sparse approximation approach.

3. The variational method used not only grants a reduction in the computational complexity, it also makes easy to incorporate input uncertainty into the model.

4. We developed a model that combines properties from both discriminative and generative approaches. This hybrid model makes possible to apply Bayesian GP-LVM on non-Gaussian data and make latent representations of mixed data types.

5. The performance of the hybrid discriminative-generative model is highly dependent on the definition of the lower dimensional representation. Better methods for improving the learning phase of this lower representation manifold still need to be researched.

6. When modelling malaria case-counts in Uganda, the data aggregation of the HMIS records precluded us from incorporating the space as an input dimension of the model. The analysis was carried on mainly with a time series approach.

However, if spatial data were available, they could be easily integrated under the modelling framework used.

7. Learning with EP-type approximations resulted considerably slow. When fitting a Poisson process we found more convenient to work on the log-space and use a GP regression model. More efficient methods for implementing EP or alternative methods for approximating such processes are needed.

8. As data collection and storage keeps improving, spatial analysis based on GIS will require the use of more efficient techniques for handling *big data*. The first step toward this direction is the assimilation of the data features into the model. We provided models for each district definition (say local in time) with a specific covariance structure and noise assumptions.

9. Another step towards the use of big data is the integration of different data sources. We were able to model jointly HMIS records and NDVI information with a vector-valued kernel. No association was found between these two sources. Nevertheless, what was not a success in terms of modelling malaria, was indeed a success in terms of showing the GP framework capabilities.

10. Perhaps one of the most important challenges within the field of spatiotemporal statistics will always be to communicate with domain-oriented sciences and planners from different sectors. We have passed the stage where we were able to provide just a mean and variance estimate, and now are able to provide density functions estimates. But these are complex outputs that need to be synthesized for different users. The monitoring system proposed is a response to such challenge.

7.1 Future Work

From the results presented in this work, we can define new routes for future work. Three main lines of research are the ones that result more appealing to the author.

Addition of spatial dimension in HMIS data modelling. The study of malaria in Uganda was constrained by the spatial resolution of the data. However, other sources of information can be used for helping understand how the disease is spreading at a local scale. Examples of these sources are population estimates, *telecom* data or mosquito maps. The addition of more data sources, as well as space as an input dimension, will make evident the need for large data methods. Quite possibly,

beyond the sparse approximations we reviewed here. Alternative methods, for this big data as well as approximations for non-Gaussian noise are needed.

Calibration and validation. The models defined in Chapter 5 were constructed based on what we could observe from the data. While cross-validation methods were used for evaluating their adequacy to the data, these methods are not useful when the underlying assumption is that the data available is biased and the error needs to be quantified. Such is the case for the harmonization method used. Remaining tasks are to harmonize the number of health facilities reporting and validating the results of the harmonization method proposed. More work has to be done to calibrate and validate the bias or error in the administrative records.

Early warning systems for disease monitoring. The monitoring system developed can be used for tracking other diseases apart form malaria. It might need modifications if applied to epidemic diseases rather than endemic ones. Also, for diseases with a clear stationary effect, periodic kernels can be used to model the seasonal component. For the moment, we only focused on the change of the short-term signal, however changes in the long-term signal also provides valuable information. Future challenges involve improving this kind of early warning systems.

References

Abrahamsen, P. (1997). *A review of Gaussian random fields and correlation functions*. Norsk Regnesentral/Norwegian Computing Center. (page 6)

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of mathematical functions*. Dover New York. (page 58)

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009a). The Gaussian process density sampler. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21, pages 9–16, Cambridge, MA. MIT Press. (page 8)

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009b). Tractable nonparametric bayesian inference in Poisson processes with Gaussian process intensities. In Bottou, L. and Littman, M., editors, *Proceedings of the International Conference in Machine Learning*, volume 26, pages 9–16, San Francisco, CA. ACM, Morgan Kaufmann. (pages 24, 106, 107, and 110)

Álvarez, M., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266. (page 13)

Álvarez, M. A. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12:1425–1466. (page 49)

Amexo, M., Tolhurst, R., Barnish, G., and Bates, I. (2004). Malaria misdiagnosis: effects on the poor and vulnerable. *The Lancet*, 364(9448):1896–1898. (page 50)

Andrade-Pacheco, R., Hensman, J., Zwießele, M., and Lawrence, N. D. (2014). Hybrid discriminative-generative approaches with Gaussian processes. In Kaski and Corander (2014). (page 7)

Andrade-Pacheco, R., Mubangizi, M., Quinn, J., and Lawrence, N. (2015). Monitoring short term changes of malaria incidence in Uganda with Gaussian processes. In Douzal-Chouakria, A. et al., editors, *Proceedings of the 1st International Workshop on Advanced Analytics and Learning on Temporal Data (AALTD)*, number 1425 in CEUR Workshop Proceedings, pages 3–9. (page 7)

Andrade-Pacheco, R., Mubangizi, M., Quinn, J., and Lawrence, N. D. (2014). Consistent mapping of government malaria records across a changing territory delimitation. *Malaria Journal*, 13(Suppl 1):P5. (page 7)

Aronszajn, P. N. (1943). La théorie des noyaux reproduisants et ses applications première partie. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 39, pages 133–153. Cambridge University Press. (page 3)

Baddeley, A., Bárány, I., and Schneider, R. (2007). Spatial point processes and their applications. *Stochastic Geometry: Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004*, pages 1–75. (page 108)

Bailey, N. T. J. (1982). *The biomathematics of malaria*. Charles Griffin & Co. Ltd. (page 49)

Baldassarre, L., Rosasco, L., Barla, A., and Verri, A. (2012). Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301. (page 12)

Barber, D. and Williams, C. K. I. (1997). Gaussian processes for Bayesian classification via hybrid Monte Carlo. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, Cambridge, MA. MIT Press. (pages 19 and 107)

Barnes, T. J. (2001). Retheorizing economic geography: from the quantitative revolution to the "cultural turn". *Annals of the Association of American Geographers*, 91(3):546–565. (page 4)

Baxter, M. and King, R. G. (1999). Measuring business cycles: approximate band-pass filters for economic time series. *Review of economics and statistics*, 81(4):575–593. (pages 2 and 82)

Bergson, H. (1889). *Essai sur les données immédiates de la conscience*. Félix Alcan. (page 1)

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kulwer Academic Publishers. (page 6)

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20. (pages 5 and 56)

Betancourt, M. and Girolami, M. (2013). Hamiltonian Monte Carlo for hierarchical models. *arXiv preprint arXiv:1312.0906*. (page 8)

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. (page 8)

Bishop, C. M. and Frey, B. J., editors (2003). *Artificial Intelligence and Statistics*, Key West, FL. (pages 100 and 101)

Blight, B. and Ott, L. (1975). A bayesian approach to model inadequacy for polynomial regression. *Biometrika*, 62(1):79–88. (page 11)

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327. (page 4)

Box, G. E. P. and Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Holden-Day. (page 3)

Burton, I. (1963). The quantitative revolution and theoretical geography. *The Canadian Geographer/Le Géographe canadien*, 7(4):151–162. (page 4)

Carter, R. and Mendis, K. N. (2002). Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews*, 15(4):564–594. (page 48)

Castellani, A. (1907). Notes on cases of fever frequently confounded with typhoid and malaria in the tropics. *Journal of Hygiene*, 7(01):1–12. (page 50)

Ceccato, P., Ghebremeskel, T., Jaiteh, M., Graves, P. M., Levy, M., Ghebreselassie, S., Ogbamariam, A., Barnston, A. G., Bell, M., del Corral, J., et al. (2007). Malaria stratification, climate, and epidemic early warning in Eritrea. *The American journal of tropical medicine and hygiene*, 77(6 Suppl):61–68. (page 73)

Chan, A. B. and Dong, D. (2011). Generalized Gaussian process models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2681–2688, Colorado Springs. (page 107)

Chatfield, C. (2013). *The analysis of time series: an introduction*. CRC Press. (page 3)

Chilès, J.-P. and Delfiner, P. (2009). *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons. (page 2)

Chu, W. and Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, pages 1019–1041. (page 20)

Cleveland, W. P. and Tiao, G. C. (1976). Decomposition of seasonal time series: A model for the census X-11 program. *Journal of the American statistical Association*, 71(355):581–587. (pages 2 and 82)

Cliff, A. D. and Ord, J. K. (1969). The problem of spatial autocorrelation. In *Studies in Regional Science (London Papers in Regional Science)*, pages 25–55. Pion. (page 5)

Cox, F. E. (2010). History of the discovery of the malaria parasites and their vectors. *Parasites and Vectors*, 3(1):5. (page 48)

Cressie, N. (1990). The origins of Kriging. *Mathematical geology*, 22(3):239–252. (page 4)

Cressie, N. (1992). *Statistics for spatial data*. Wiley-Interscience. (page 5)

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons. (pages 1 and 5)

Damianou, A., Ek, C. H., Titsias, M. K., and Lawrence, N. D. (2012). Manifold relevance determination. In Langford, J. and Pineau, J., editors, *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA. Morgan Kaufmann. (pages 36 and 43)

Damianou, A. and Lawrence, N. D. (2013). Deep Gaussian processes. In Carvalho, C. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA. JMLR W&CP 31. (page 46)

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473. (page 3)

Diggle, P. J. (2003). *Statistical analysis of spatial point patterns.* Edward Arnold. (pages 24 and 108)

Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M., et al. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563. (page 5)

Diggle, P. J., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350. (page 5)

Durrande, N., Hensman, J., Rattray, M., and Lawrence, N. D. (2013). Gaussian process models for periodicity detection. *arXiv preprint arXiv:1303.7090*. (page 82)

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, pages 987–1007. (page 4)

Escalante, A. A., Barrio, E., and Ayala, F. J. (1995). Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. *Molecular biology and evolution*, 12(4):616–626. (page 48)

Fahrmeir, L., Tutz, G., Hennevogl, W., and Salem, E. (1994). *Multivariate statistical modelling based on generalized linear models*, volume 2. Springer New York. (page 107)

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd. (page 4)

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, pages 246–263. (page 5)

Gandin, L. S. (1963). *Ob"ektivnyi analiz meteorologicheskikh polei*. Gidrometeologicheskoe Izdatel'stvo, Leningrad. Translation (1965): *Objective analysis of meteorological fields*. Israel Program for Scientific Translations, Jerusalem. (page 4)

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*. CRC Press, 3th edition. (pages 58 and 61)

Gelman, A., Hwang, J., and Vehtari, A. (2013). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, pages 1–20. (page 11)

Ghahramani, Z., editor (2007). *Proceedings of the International Conference in Machine Learning*, volume 24. Omnipress. (page 102)

Gibbs, M. N. and MacKay, D. J. C. (2000). Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464. (pages 19 and 107)

Girard, A. and Murray-Smith, R. (2005). Gaussian processes: Prediction at a noisy input and application to iterative multiple-step ahead forecasting of time-series. In Murray-Smith, R. and Shorten, R., editors, *Switching and Learning in Feedback Systems*, Lecture Notes in Computer Science, pages 158–184. Springer. (page 33)

Girard, A., Rasmussen, C. E., Quiñonero Candela, J., and Murray-Smith, R. (2003). Gaussian process priors with uncertain inputs – application to multiple-step ahead time series forecasting. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15, pages 529–536, Cambridge, MA. MIT Press. (page 33)

Goldberg, P. W., Williams, C. K. I., and Bishop, C. M. (1998). Regression with input-dependent noise: A Gaussian process treatment. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*, volume 10, pages 493–499, Cambridge, MA. MIT Press. (page 65)

Good, M. F. and Doolan, D. L. (1999). Immune effector mechanisms in malaria. *Current opinion in immunology*, 11(4):412–419. (page 49)

Gosoniu, L., Vounatsou, P., Sogoba, N., and Smith, T. (2006). Bayesian modelling of geostatistical malaria risk data. *Geospatial Health 1*, pages 127–139. (page 49)

Grigoriu, M. (2002). *Stochastic calculus: applications in science and engineering*. Springer. (page 3)

Haque, U., Hashizume, M., Glass, G. E., Dewan, A. M., Overgaard, H. J., and Yamamoto, T. (2010). The role of climate variability in the spread of malaria in Bangladeshi highlands. *PLoS One*, 5(12):e14341. (page 73)

Hay, S. I., Snow, R. W., and Rogers, D. J. (1998). From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives. *Parasitology Today*, 14(8):306–313. (page 84)

Hein, M. and Bousquet, O. (2004). Kernels, associated structures and generalizations. Technical report, Max Planck Institute for Biological Cybernetics. (pages 6 and 13)

Helterbrand, J. D. and Cressie, N. (1994). Universal cokriging under intrinsic coregionalization. *Mathematical Geology*, 26(2):205–226. (page 13)

Hensman, J., Rattray, M., and Lawrence, N. D. (2012). Fast variational inference in the conjugate exponential family. In Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25, Cambridge, MA. (page 35)

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*. Siam. (page 14)

Hinton, G. E. and van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. (pages 8 and 104)

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430. (pages 2 and 82)

Jaakkola, T. S. and Jordan, M. I. (1996). Computing upper and lower bounds on likelihoods in intractable networks. In Horvitz, E. and Jensen, F. V., editors, *Uncertainty in Artificial Intelligence*, volume 12, San Francisco, CA. Morgan Kaufmann. (pages 19 and 34)

Joy, D. A., Feng, X., Mu, J., Furuya, T., Chotivanich, K., Krettli, A. U., Ho, M., Wang, A., White, N. J., Suh, E., et al. (2003). Early origin and recent expansion of Plasmodium falciparum. *Science*, 300(5617):318–321. (page 48)

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011). Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12:3227–3257. (page 20)

Karhunen, K. (1947). Über lineare methoden in der wahrscheinlichkeitsrechnung. *Annales Academie Scientiarum Fennicae*, A.I. 37. (pages 3 and 6)

Kaski, S. and Corander, J., editors (2014). *Artificial Intelligence and Statistics*, volume 33, Iceland. JMLR W&CP. (page 92)

Kearns, M. J., Solla, S. A., and Cohn, D. A., editors (1999). *Advances in Neural Information Processing Systems*, volume 11, Cambridge, MA. MIT Press. (pages 98 and 101)

Kim, A., Thomas, R., Aldering, G., Antilogus, P., Aragon, C., Bailey, S., Baltay, C., Bongard, S., Buton, C., Canto, A., et al. (2013). Standardizing type Ia supernova absolute magnitudes using Gaussian process data regression. *The Astrophysical Journal*, 766(2):84. (page 11)

Kintu, P., Nanyunja, M., Nzabanita, A., and Magoola, R. (2005). Development of HMIS in poor countries: Uganda as a case study. *Health Policy and Development*, 3(1):46–53. (page 50)

Kleinschmidt, I., Bagayoko, M., Clarke, G. P. Y., Craig, M., and Sueur, D. L. (2000). A spatial statistical approach to malaria mapping. *International Journal of Epidemiology*, 29(2):355–361. (page 49)

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614. (page 8)

Kolmogorov, A. N. (1941). Interpolation und extrapolation von stationären zufälligen folgen. *Izvestiia Akademii Nauk SSSR, Seriia Matematicheskai*, 5(1):3–14. Translation (1962): Interpolation and extrapolation of stationary random sequences. Rand Corporation, California. (page 4)

Korattikara, A., Chen, Y., and Welling, M. (2013). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. *arXiv preprint arXiv:1304.5299*. (page 8)

Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704. (pages 19, 20, 21, 25, and 104)

Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, USA. (page 36)

Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816. (page 35)

Lawrence, N. D. (2007). Learning for larger datasets with the Gaussian process latent variable model. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics*, pages 243–250, San Juan, Puerto Rico. Omnipress. (page 14)

Lawrence, N. D. and Jordan, M. I. (2005). Semi-supervised learning via Gaussian processes. In Saul, L., Weiss, Y., and Bouttou, L., editors, *Advances in Neural Information Processing Systems*, volume 17, pages 753–760, Cambridge, MA. MIT Press. (page 20)

Lázaro-Gredilla, M. and Titsias, M. (2011). Variational heteroscedastic Gaussian process regression. In Getoor, L. and Scheffer, T., editors, *Proceedings of the International Conference in Machine Learning*, volume 28, pages 841–848, San Francisco, CA. Morgan Kaufmann. (page 65)

Lee, D. D. and Sompolinsky, H. (1999). Learning a continuous hidden variable model for binary data. In Kearns et al. (1999). (page 34)

Liu, W., Li, Y., Learn, G. H., Rudicell, R. S., Robertson, J. D., Keele, B. F., Ndjango, J.-B. N., Sanz, C. M., Morgan, D. B., Locatelli, S., et al. (2010). Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature*, 467(7314):420–425. (page 48)

Loève, M. (1948). *Fonctions aléatoires du second ordre*. Gauthier-Villars, Paris. Supplement to P. Lévy, *Processus stochastiques et mouvement Brownien*. (pages 3 and 6)

Manh, B. H., Clements, A. C., Thieu, N. Q., Hung, N. M., Hung, L. X., Hay, S. I., Hien, T. T., Wertheim, H. F., Snow, R. W., and Horby, P. (2011). Social and environmental determinants of malaria in space and time in Viet Nam. *International journal for parasitology*, 41(1):109–116. (page 75)

Masani, P. (1966). Wiener's contributions to generalized harmonic analysis, prediction theory and filter theory. *Bulletin of the American Mathematical Society*, 72:73–125. (page 4)

Matheron, G. (1962). Traité de géostatistique appliquée, tome I. *Mémoires du Bureau de Recherche Géologiques et Minières*, No. 14. Editions Technip, Paris. (page 4)

Matheron, G. (1963). Traité de géostatistique appliquée, tome II: le krigeage. *Mémoires du Bureau de Recherche Géologiques et Minières*, No. 24. Editions Bureau de Recherche Géologiques et Minières, Paris. (page 4)

Matheron, G. (1982). Pour une analyse krigeante de donnés régionalisées. Technical report, École des Mines de Paris, Fontainebleau, France. (page 12)

Micchelli, C. A. and Pontil, M. (2004). Kernels for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press. (page 12)

Micchelli, C. A. and Pontil, M. (2005). On learning vector–valued functions. *Neural Computation*, 17:177–204. (page 12)

Ministry of Health, Health Systems 20/20, and Makerere University School of Public Health (2012). Uganda health system assessment 2011. Technical report, MD: Health Systems 20/20 project, Kampala, Uganda and Bethesda. (page 48)

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. In Breese, J. S. and Koller, D., editors, *Uncertainty in Artificial Intelligence*, volume 17, pages 362–369, San Francisco, CA. Morgan Kaufmann. (pages 9, 20, and 34)

Møller, J. and Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684. (pages 24, 109, and 110)

Mubangizi, M., Andrade-Pacheco, R., Smith, M., Quinn, J. A., and Lawrence, N. (2014). Malaria surveillance with multiple data sources using Gaussian process models. In *1st International Conference on the Use of Mobile ICT in Africa 2014*. (page 7)

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press. (page 1)

Myers, D. E. (1982). Matrix formulation of co-Kriging. *Journal of the International Association for Mathematical Geology*, 14(3):249–257. (page 12)

Naish-Guzman, A. and Holden, S. (2008). The generalized FITC approximation. *Advances in Neural Information Processing Systems*, 20:1057–1064. (pages 22, 27, and 37)

Neal, R. M. (1998). Regression and classification using Gaussian process priors. *Bayesian Statistics*, 6:475–502. (pages 8 and 106)

Nelder, J. A. and Wedderburn, W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, 135(3):370–384. (page 106)

O'Hagan, A. and Kingman, J. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–42. (page 11)

Opper, M. and Winther, O. (2000). Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655–2684. (pages 20 and 34)

Palmer, J., Kreutz-Delgado, K., Rao, B. D., and Wipf, D. P. (2005). Variational EM algorithms for non-Gaussian latent variable models. In Weiss et al. (2006), pages 1059–1066. (pages 9 and 104)

Parzen, E. (1959). Statistical inference on time series by Hilbert space methods. Technical Report 23, Stanford University. (pages 3 and 6)

Parzen, E. (1961). An approach to time series analysis. *The Annals of Mathematical Statistics*, pages 951–989. (page 3)

Parzen, E. (1970). Statistical inference on time series by RKHS methods. In Pyke, R., editor, *12th Biennial Seminar*, pages 1–37. Canadian Mathematical Congress. (page 3)

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, pages 25–45. (page 5)

Portugal, S., Carret, C., Recker, M., Armitage, A. E., Gonçalves, L. A., Epiphanio, S., Sullivan, D., Roy, C., Newbold, C. I., Drakesmith, H., et al. (2011). Host-mediated regulation of superinfection in malaria. *Nature medicine*, 17(6):732–737. (page 49)

Quenouille, H. (1957). *The analysis of multiple time-series*. Griffin's statistical monographs & courses. Griffin. (pages 3 and 12)

Quiñonero Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959. (page 14)

Quinn, J. A., Leyton-Brown, K., and Mwebaze, E. (2011). Modeling and monitoring crop disease in developing countries. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*. (page 49)

Reinerand, R. C., Perkins, T. A., Barker, C. M., Niu, T., Chaves, L. F., Ellis, A. M., George, D. B., Menach, A. L., Pulliam, J. R. C., Bisanzio, D., et al. (2013). A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. *Journal of The Royal Society Interface*, 10(81):20120921. (page 49)

Riesz, F. and Sz-Nagy, B. (1955). *Functional Analysis*. Frederick Ungar, New York. (page 6)

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, B*, 71(2):319–392. (pages 9 and 104)

Särkkä, S. (2011). Linear operators and stochastic partial differential equations in Gaussian process regression. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 151–158. Springer. (page 83)

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

Schein, A. I., Saul, L. K., and Ungar, L. H. (2003). A generalized linear model for principal component analysis of binary data. In Bishop and Frey (2003). (page 34)

Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism*, 3(1):13–41. (page 2)

Seeger, M. (2004). Gaussian processes for Machine Learning. *International Journal of Neural Systems*, 14(2):69–106. (pages 19 and 107)

Seeger, M. (2005). Expectation propagation for exponential families. Technical report, University of California at Berkeley. (pages 9 and 20)

Seeger, M., Williams, C. K. I., and Lawrence, N. D. (2003). Fast forward selection to speed up sparse Gaussian process regression. In Bishop and Frey (2003). (page 14)

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K. (page 6)

Slutzky, E. (1927). Slozhenie sluchainykh prichin, kak istochnik tsiklicheskikh protsessov'. *Voprosy kon"yunktury*, 1(3):34–64. Translation (1937): The summation of random causes as the source of cyclic processes. *Econometrica*. (pages 2 and 5)

Smith, D. L., Battle, K. E., Hay, S. I., Barker, C. M., Scott, T. W., and McKenzie, F. E. (2012). Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS pathogens*, 8(4):e1002588. (page 49)

Snelson, E. and Ghahramani, Z. (2006a). Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 18:1257–1264. (pages 8 and 27)

Snelson, E. and Ghahramani, Z. (2006b). Sparse Gaussian processes using pseudo-inputs. In Weiss et al. (2006). (page 14)

Stern, D. I., Gething, P. W., Kabaria, C. W., Temperley, W. H., Noor, A. M., Okiro, E. A., Shanks, G. D., Snow, R. W., and Hay, S. I. (2011). Temperature and malaria trends in highland East Africa. *PLoS One*, 6(9):e24524. (page 75)

Sudakov, V. N. (1993). Gaussian measures. A brief survey. In *Workshop di Teoria della Misura e Analisi Reale*, Grado, Italia. (page 6)

Tipping, M. E. (1999). Probabilistic visualisation of high-dimensional binary data. In Kearns et al. (1999), pages 592–598. (page 34)

Tipping, M. E. and Lawrence, N. D. (2003). A variational approach to robust Bayesian interpolation. In Molina, C., Adali, T., Larsen, J., Hulle, M. V., Douglas, S., and Rouat, J., editors, *Neural Networks for Signal Processing XIII*, pages 229–238. IEEE. (page 19)

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL. JMLR W&CP. (pages 14, 18, 24, 28, and 34)

Titsias, M. K. and Lawrence, N. D. (2010). Bayesian Gaussian process latent variable model. In Teh, Y. W. and Titterington, D. M., editors, *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics*, volume 9, pages 844–851, Chia Laguna Resort, Sardinia, Italy. JMLR W&CP. (pages 34, 35, and 36)

Tolvanen, V., Jylanki, P., and Vehtari, A. (2014). Expectation propagation for nonstationary heteroscedastic Gaussian process regression. In *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pages 1–6. IEEE. (page 65)

Trigg, P. I. and Kondrachine, A. V. (1998). Commentary: malaria control in the 1990s. *Bulletin of the World Health Organization*, 76(1):11. (page 48)

Urtasun, R. and Darrell, T. (2007). Discriminative Gaussian process latent variable model for classification. In Ghahramani (2007). (pages 34, 44, and 46)

van Ruth, F., Schouten, B., and Wekker, R. (2005). The statistics Netherlands' business cycle tracer. Methodological aspects; concept, cycle computation and indicator selection. Technical report, Statistics Netherlands. (page 83)

Vanhatalo, J. (2006). Sparse log Gaussian process in spatial epidemiology. Master's thesis, Helsinki University of Technology. (page 49)

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in medicine*, 29(15):1580–1607. (pages 25 and 104)

Vehtari, A., Ojanen, J., et al. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228. (pages 57 and 111)

Vehtari, A., Tolvanen, V., Mononen, T., and Winther, O. (2014). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *arXiv preprint arXiv:1412.7461*. (pages 57 and 112)

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., and von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172. (page 1)

Weiss, Y., Schölkopf, B., and Platt, J. C., editors (2006). *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA. MIT Press. (pages 99 and 101)

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449. (page 4)

Wiener, N. (1942). *Extrapolation, interpolation, and smoothing of stationary time series*. Report of the Services 19. MIT Press. Printed in book form (1949): John Wiley & Sons, New York. (page 4)

Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351. (pages 9, 104, and 106)

Williams, C. K. I. and Rasmussen, C. E. (2006). *Gaussian processes for Machine Learning*. MIT Press. (pages 20, 57, and 111)

Wold, H. O. (1938). *A Study in the Analysis of Stationary Time Series*. A Study in the Analysis of Stationary Time Series. Almquist & Wiksell, Stockholm. (pages 3 and 4)

World Health Organization (2010). Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. Technical report, WHO Press, Geneva. (page 49)

World Health Organization (2011). Country health information systems: a review of the current situation and trends. Technical report, WHO Press, Geneva. (page 49)

World Health Organization (2015). World health statistics 2015. Technical report, WHO Press, Geneva. (page 48)

World Health Organization and others (2014). World malaria report 2014. Technical report, WHO Press, Geneva. (page 48)

World Health Organization and Uganda Ministry of Health (2011). Assessment of health facility data quality. Data quality report card Uganda, 2010-2011. Technical report, WHO Press, Geneva. (page 50)

Wu, Y., Hernández-Lobato, J. M., and Ghahramani, Z. (2014). Gaussian process volatility model. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27, pages 1044–1052, Cambridge, MA. (page 65)

Yaglom, A. M. (1955). The correlation of processes whose nth differences constitute a stationary process. *Mat. Sb.*, 37(79):141. (page 3)

Yaglom, A. M. (1986a). *Correlation theory of stationary and related random functions I: Basic results*. Springer-Verlang. (page 12)

Yaglom, A. M. (1986b). *Correlation theory of stationary and related random functions II: Supplementary notes and references*. Springer-Verlang. (page 12)

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pages 267–298. (page 2)

Appendix A

Alternative Methods for Approximate Inference

In this work we have chosen to use expectation propagation as approximation method and explore ways to extend it. However, as discussed in Chapter 2, there are alternative approximation methods. We are interested in EP, as it has proved to be effective in some cases (Kuss and Rasmussen, 2005; Vanhatalo et al., 2010). Nevertheless, there might be good reasons to use a different method or combine them together. For instance, the variational framework we use, works under the principles of variational Bayes (Hinton and van Camp, 1993; Palmer et al., 2005).

Given the importance of the variational approach to our research, we briefly present here the core ideas behind the variational Bayes approximation. Also, we are aware that within the spatial statistics community, Laplace approximation (or its evolution into INLA) has received a lot of attention in recent years (Rue et al., 2009; Williams and Barber, 1998). For this reason, we consider important to present this approximation here as well.

A.1 Variational Bayes

Variational Bayes consists of approximating an intractable posterior distribution $p(Z|X)$ with a distribution $q(Z)$. This approximation is obtained by minimizing $\mathrm{KL}\,(q\|p)$. Tractability of $q(Z)$ is ensured by restricting it to a specific family of distributions, for example: a parametric distribution $q(Z|\omega)$ governed by a set of parameters ω; or a family of distributions that can be factorized as $q(Z) = \prod q_i(Z_i)$, where Z_i are disjoint groups of the elements of Z.

In general, KL minimization relies in the fact that any distribution $p(X)$ can be decomposed as

$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}\left(q \| p\right), \tag{A.1}$$

where

$$\mathcal{L}(q) = \int q(Z) \log \frac{p(X,Z)}{q(Z)} \mathrm{d}Z, \tag{A.2}$$

$$\mathrm{KL}\left(q \| p\right) = -\int q(Z) \log \frac{p(Z|X)}{q(Z)} \mathrm{d}Z. \tag{A.3}$$

Thus minimizing $\mathrm{KL}\left(q \| p\right)$ is equivalent to maximizing $\mathcal{L}(q)$.

A.2 Laplace Approximation

The Laplace approximation consists of approximating a posterior distribution with a Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{A})$. The parameters of the approximation are defined according to the second order Taylor expansion over the logarithm (logit) of the posterior distribution.

Let $p(\mathbf{f}|\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})$, with $p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$, then

$$\begin{aligned} \ln p(\mathbf{f}|\mathbf{y}) &= \ln p(\mathbf{y}|\mathbf{f}) + \ln p(\mathbf{f}|\mathbf{X}) \\ &= \ln p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\ln|\mathbf{K}| - \frac{n}{2}\ln 2\pi. \end{aligned} \tag{A.4}$$

Using the second order Taylor expansion of the log-posterior around its mode $\mathbf{m}$, we have that

$$\ln p(\mathbf{f}|\mathbf{y}) \approx \ln p(\mathbf{m}|\mathbf{f}) - \frac{1}{2}(\mathbf{m}-\mathbf{f})^\top \mathbf{A}^{-1}(\mathbf{m}-\mathbf{f}), \tag{A.5}$$

where

$$\mathbf{m} = \underset{\mathbf{f}}{\operatorname{argmax}}\, p(\mathbf{f}|\mathbf{y}), \tag{A.6}$$

$$\mathbf{A} = -\nabla\nabla \ln p(\mathbf{f}|\mathbf{y})|_{\mathbf{f}=\mathbf{m}}. \tag{A.7}$$

Appendix B

Generalized Linear Models and Gaussian Processes

When modelling spatiotemporal processes, it is usually necessary to decide on the adequacy and implications of using non-Gaussian noise models. Despite the flexibility of Gaussian processes, their use in a regression model implies assuming symmetry, continuity and no bounds in the range of values of the output variable. This can be too restrictive or not realistic for some applications. A more involved model than the standard regression, able to guarantee a wider range of assumptions in the values of the output, can be defined by using a GP as a latent process embedded in a non-Gaussian process (see for example Adams et al. (2009b); Neal (1998); Williams and Barber (1998)). For a set of input-output observations $\{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathbb{R}^q, i = 1, \ldots, n\}$ if the output y_i is not Gaussian, its likelihood can be modelled as

$$p(y_i|f_{\mathbf{x}_i}) = g^{-1}(f_{\mathbf{x}_i}), \tag{B.1}$$

where $g^{-1}(\cdot)$ is a monotonic differentiable function and $(f_{\mathbf{x}_i}) \sim \mathcal{GP}$. This approach was introduced by Nelder and Wedderburn (1972) in a parametric formulation, known as *generalized linear models* (GLM).

B.1 GLM Formulation

A GLM is characterized for having a linear predictor

$$\eta_i = \mathbf{w}_i^\top \mathbf{x}_i, \tag{B.2}$$

for some $\mathbf{w}_i \in \mathbb{R}^q$, such that $\exists$ a monotonic differentiable function $g(\cdot)$ that links the linear predictor η_i with the mean of the process y_i through the relation

$$g(\langle y_i \rangle) = \eta_i. \tag{B.3}$$

The election of the link function is not uniquely defined by the distribution assumed for y_i, however its common that some distributions are assigned a particular link function (Fahrmeir et al., 1994). For example, the Gaussian distribution is frequently used with the *identity* function and the Poisson distribution is frequently used with a *logarithmic* transformation.

A GLM can be implemented in a non-parametric setting with the aid of a latent GP by letting

$$g(\langle y_i \rangle) = f_{\mathbf{x}_i}, \tag{B.4}$$

where $(f_{\mathbf{x}_i}) \sim \mathcal{GP}$. This model keeps the flexibility of the Gaussian processes and at the same time incorporates prior knowledge about the relation between the process mean and the latent variable (Chan and Dong, 2011). Bayesian inference on Equation (B.4) requires the computation of the posterior distribution

$$p(f_{\mathbf{x}_i}|y_i) = \frac{p(f_{\mathbf{x}_i})p(y_i|f_{\mathbf{x}_i})}{p(y_i)}, \tag{B.5}$$

where $p(y_i|f_{\mathbf{x}_i})$ is given by Equation (B.1). Notice that dependence on $\mathbf{x}_i$, in Equations (B.1) and (B.5) has been omitted to simplify notation. Transformation (B.1) usually makes the posterior distribution intractable, and approximations are needed (Adams et al., 2009b; Barber and Williams, 1997; Gibbs and MacKay, 2000; Seeger, 2004).

Appendix C

Point Processes

Point pattern data is so common across a wide variety of fields that it would be surprising if a textbook on spatial statistics did not dedicate a section to its study. Beyond the use of Poisson likelihoods and binary classification problems, we do not need to delve much into point processes along this thesis. However, we define them formally here for completeness.

Within stochastic processes literature, Poisson processes are generally the first stop when moving from the continuous to the discrete case. Here we present a brief introduction and explain how they can implemented with the aid of a Gaussian process.

C.1 Point Processes Formulation

A point process is a stochastic process characterized by generating a countable set of events $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ across a region $\mathbb{S}$ (Diggle, 2003). In general, the study of these processes is based on the observed locations of an event of interest, within a bounded region. When the space $\mathbb{S}$ is unidimensional, say time for instance, point processes can be handled in three different ways (Baddeley et al., 2007): they can be studied through the *arrival times* $\{T_i, T_2, \ldots | T_1 < T_2 < \ldots\}$, where T_i is the time at which event $\mathbf{x}_i$ occurs; or through the *inter-arrival times* $\{W_i = T_{i+1} - T_i\}$, where W_i is the time between two consecutive events; or through a *cumulative counting process* $\mathring{n}(t) = \sum_{i=1}^{\infty} \mathcal{I}_{\{T_i<t\}}$, where $\mathcal{I}$ is an index function such that $\mathcal{I} : \mathbb{R} \to \{0, 1\}$. Due to there is no natural order in higher dimensional spaces, there is no equivalent for inter-arrival times or cumulative counting process in that case. The alternative for handling point processes in dimensions greater than one is through region counts $n_{\mathbb{B}}$, where $\mathbb{B} \subset \mathbb{S}$ is a bounded set.

Definition 7 *Let $\mathbb{S}$ be a complete, separable, metric space, and let $\{\mathbb{B}_1, \ldots, \mathbb{B}_m | \mathbb{B}_i \subset \mathbb{S}, m \in \mathbb{N}\}$ be a collection of bounded Borel sets. A point process is a collection of non-negative random variables $\{n_{\mathbb{B}_1}, \ldots, n_{\mathbb{B}_m}\}$, so that the points of the process are those locations $\mathbf{x}_i \in \mathbb{S}$ where $n_{\{\mathbf{x}_i\}} > 0$.*

It is usually assumed that any bounded region $\mathbb{B}_i$ contains only a finite number of points with probability 1. This is

$$p(n_{\mathbb{B}_i} < \infty) = 1. \tag{C.1}$$

If it is also assumed that no two events are coincident, i.e.,

$$n_{\{\mathbf{x}_i\}} \leq 1, \forall \mathbf{x}_i \in \mathbb{S}, \tag{C.2}$$

the process is known as *simple point process*. When the locations $\mathbf{x}_i$ are associated to a random variable $m_{\mathbf{x}_i}$, which provides more information about the event, the process is known as a *marked point process* (Møller and Waagepetersen, 2007).

C.2 Poisson Process

Definition 8 *Let $\lambda : \mathbb{S} \to [0, \infty)$ be an intensity function that is locally integrable (i.e., $\int_{\mathbb{B}} \lambda(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty$ for all bounded $\mathbb{B} \subseteq \mathbb{S}$), and let $\mu(\mathbb{B}) = \int_{\mathbb{B}} \lambda(\mathbb{B}) \mathrm{d}\mathbf{x}$ be an intensity measure, such that it is locally finite (i.e., $\mu(\mathbb{B}) < \infty$ for bounded $\mathbb{B} \subseteq \mathbb{S}$) and diffuse (i.e., $\mu(\{\mathbf{x}\}) = 0$, for all $\mathbf{x} \in \mathbb{S}$). Then, a Poisson Process $(n_{\mathbb{B}})_{\mathbb{B} \subset \mathbb{S}}$ defined on $\mathbb{S}$ with intensity measure $\mu(\cdot)$ and intensity function $\lambda(\cdot)$ is a point process that satisfies the following conditions:*

1. $n_{\mathbb{B}} \sim \text{Poisson}\,(\mu(\mathbb{B}))$.
2. *Conditional on $n_{\mathbb{B}}$, the points $\mathbf{x}_i \in \mathbf{X}_{\mathbb{B}}$ are iid. with density proportional to $\lambda(\mathbf{x}_i)$.*

We will denote a Poisson process as $(n_{\mathbb{B}})_{\mathbb{B} \subset \mathbb{S}} \sim \mathcal{PP}(\lambda)$, where λ is an intensity function.

If the intensity function is constant: $\lambda(s) = \lambda \in \mathbb{R}^+$, the Poisson process is said to be *homogeneous*, otherwise it is said to be *inhomogeneous*. In the former, the probability of observing any point pattern does not depend on the location of its points. Although the homogeneous process is easy to interpret and is analytically tractable in Bayesian computation, its assumptions are too restrictive or unrealistic for many

applications (Møller and Waagepetersen, 2007). The inhomogeneous Poisson process, although analytically intractable (Adams et al., 2009b), is a less restrictive alternative in terms of modelling. Different models arise depending on the way the intensity function is modelled. Here we present three variants:

- **Cox Process:** Let $\lambda(\mathbf{x})$, with $\mathbf{x} \in \mathbb{S}$, be a non-negative process. A Cox Process is a Point Process where $n_{\mathbb{B}}|\lambda(\mathbf{x}) \sim \text{Poisson}(\mu(\mathbb{B}))$. The difference between this definition and the one given for a Poisson process above is that here $\lambda(\mathbf{x})$ is also stochastic.

- **Log-Gaussian Cox Process:** We call $(n_{\mathbb{B}}) \sim \mathcal{PP}(\lambda)$ a log-Gaussian Cox process if the intensity function is defined as $\log \lambda(\mathbf{x}) = \boldsymbol{\beta}^{\intercal}\phi(\mathbf{x}) + f_{\mathbf{x}}$; where $\boldsymbol{\beta}$ is a parameter, $\phi(\cdot)$ is a basis function and $(f_{\mathbf{x}}) \sim \mathcal{GP}(\mathcal{M}, K)$.

- **Sigmoidal Gaussian Cox Process:** We call $(n_{\mathbb{B}}) \sim \mathcal{PP}(\lambda)$ a sigmoidal Gaussian Cox process if the intensity function is given by $\lambda(\mathbf{x}) = \lambda_* \sigma(f_{\mathbf{x}})$, with $\lambda_* \in \mathbb{R}^+$, $\sigma(z) = (1 + e^{-z})^{-1}$ and $(f_{\mathbf{x}}) \sim \mathcal{GP}(\mathcal{M}, K)$.

Appendix D

Model Validation

Once a learning model has been defined and trained, it is needed to assess its *goodness* to know how adequate the model is and how reliable its results are. Among the several methods that can be used for this purpose (see Vehtari et al. (2012)), hold-out sample methods provide a simple and intuitive procedure. These methods consists of splitting the data set into two disjoint sets, one used for training or fitting the model and the other for evaluating its performance. A clear drawback of this procedure is that not all data is being used during the learning phase. An alternative implementation of hold-out sample validation that uses all data for training the model is k-fold cross-validation.

D.1 Cross-Validation

In k-fold cross-validation, data is split into k disjoint sets, ideally all of them of the same size. The model is fitted k times, each time trained with the data contained in $k-1$ sets and validated with the remaining set. All data points are used for learning, although not at the same time. The cost of implementing this validation procedure involves training k models, which can be prohibitive in some cases.

The extreme case of k-fold cross-validation is given when the number of disjoint sets equals the number of data points. This means that the validation sets consist of a single observation that was left out of the training. This case of validation is know as leave-one-out cross-validation (LOO-CV). Although it can look like the most expensive case of cross-validation, in the case of Gaussian process models, there are almost no additional computations beyond the learning phase (Williams and Rasmussen, 2006).

For a set of observations $\{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{y} = (y_1 \dots, y_n)^\top$ and $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, the log-predictive probability of of y_i, when the model was trained only with the

observations in $\mathbf{y}_{-i} = \{y_j | j \neq i\}$ is given by

$$\log p(y_i|\mathbf{y}_{-i}, \mathbf{X}) = -\frac{1}{2}\log 2\pi\sigma_i^2 - \frac{(y_i - \mu_i)^2}{2\sigma_i^2}. \tag{D.1}$$

In a Gaussian process regression, the parameters μ_i and σ_i are computed as

$$\mu_i = y_i - \frac{[(\mathbf{K_{ff}} + \sigma^2\mathbf{I})^{-1}\mathbf{y}]_{ii}}{[(\mathbf{K_{ff}} + \sigma^2\mathbf{I})^{-1}]_{ii}}, \tag{D.2}$$

$$\sigma_i^2 = \frac{1}{[(\mathbf{K_{ff}} + \sigma^2\mathbf{I})^{-1}]_{ii}}. \tag{D.3}$$

The terms involved in the computation of μ_i and σ_i^2 are all computed during the training of the model. This is why there is not additional cost in computing LOO-CV. The predictive log-probability of the whole set is then

$$Ł(\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{-i}, \mathbf{X}). \tag{D.4}$$

When using expectation propagation, LOO-CV cannot be computed using equations (D.2) and (D.3). But still no additional computations are needed. The cavity distribution $q_{-i}(f_{\mathbf{x}_i})$, computed for each data point, is the leave-one-out marginal posterior of the latent variable. Hence, the approximate LOO-CV predictive probability can be computed as

$$p(y_i|\mathbf{y}_{-i}, \mathbf{X}) = \int q_{-i}(f_{\mathbf{x}_i}) p(y_i|f_{\mathbf{x}_i}) \mathrm{d}f_{\mathbf{x}_i}, \tag{D.5}$$

which is the zero-th moment of the factor approximation $\ell_i(f_{\mathbf{x}_i})$, and is already computed when training the model (Vehtari et al., 2014).

In the sparse approximations presented in this thesis, it is not clear how to define a shortcut for computing LOO-CV, as the matrix $\mathbf{Q_{ff}} = \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}$, used to approximate the covariance matrix, encodes information from all data points.

Appendix E

Results

E.1 Noise Model Selection

Table E.1 Noise model selection. The second column shows the number of data points used in each district. The third and fourth columns show the LOO-CV log-predictive probabilities of the Gaussian and log Gaussian models.

District	N	Gaussian	log Gaussian
Adjumani	528	-374.8	-189.9
Busia	459	80.3	-108.2
Hoima	556	-485.7	-256.9
Jinja	555	65.2	67.9
Kabale	571	-473.8	-306.0
Kabarole	562	-458.9	-260.7
Kaberamaido	512	-538.5	-336.4
Kalangala	433	-543.8	-215.6
Kampala	361	48.5	29.4
Kamwenge	559	-835.2	-323.0
Kanungu	485	9.6	-153.7
Kasese	489	-120.1	-224.4
Kayunga	565	103.8	35.4
Kibaale	449	-546.7	-144.0
Kisoro	565	-194.6	42.1
Mayuge	439	-459.7	-251.2

Continued on next page

District	N	Gaussian	log Gaussian
Moyo	578	-348.6	-384.6
Nakasongola	487	-495.1	-117.9
Ntungamo	481	-565.7	-181.3
Rukungiri	544	-192.2	-149.3
Ssembabule	578	-313.3	-136.4
Wakiso	561	-416.3	-349.2
Yumbe	499	-545.2	-172.0

E.2 Kernel Selection

E.2.1 RBF vs Matérn-3/2

Table E.2 Kernel selection for time dependence. The consecutive numbers at the end of the district names indicate a change in the districts delimitation. The second column shows the number of data points en each case. The third and fourth columns show the log LOO-CV scores for the RBF and Mat-3/2 kernels. When the number of observations available was less than 10, no comparison was done and a Mat-3/2 kernel was used.

District	N	RBF	Mat-32
Alebtong	181	-102.6	-105.1
Bundibugyo-1	381	-140.8	-53.9
Bundibugyo-2	190	-67.0	-31.6
Buyende	156	-63.4	-62.0
Gomba	148	-85.4	-73.4
Jinja	555	-28.9	122.1
Kaberamaido	512	-279.2	-278.5
Kalangala	433	-193.5	-194.7
Kaliro	376	-209.5	-206.8
Kasese	489	-220.7	-209.1
Katakwi-1	145	-28.4	-26.9
Katakwi-2	295	-175.3	-171.6
Mpigi-1	349	-122.9	-121.7
Mpigi-3	163	-15.4	-15.5

Continued on next page

District	N	RBF	Mat-32
Serere	120	-39.1	-37.0
Yumbe	499	-246.6	-142.6
Kayunga	565	44.4	50.0
Bushenyi-1	386	-72.6	-54.4
Bushenyi-4	24	6.5	8.4
Bushenyi-5	162	-76.7	-75.3
Gulu-1	184	-101.0	-89.2
Gulu-2	198	4.2	4.7
Gulu-3	158	-133.5	-16.5
Kampala	565	-354.8	-347.2
Namayingo	98	-138.7	-139.0
Lyantonde	248	-228.7	-193.6
Nwoya	180	-114.3	-101.8
Pader-1	374	-221.2	-109.7
Pader-2	135	-89.5	-87.4
Soroti-1	361	-248.3	-233.4
Soroti-2	143	-54.8	-52.9
Abim	227	-141.9	-100.4
Arua-1	162	29.5	29.1
Arua-2	62	33.1	33.1
Arua-3	307	-217.0	-207.0
Kabarole	562	-246.3	-244.1
Mityana	391	-70.5	-69.8
Sironko-1	340	-39.3	-38.7
Sironko-2	149	-61.7	-60.9
Bukwo	396	-221.9	-220.1
Busia	459	-141.3	-85.5
Butaleja	407	104.0	112.0
Isingiro	307	-122.3	-118.6
Kyegegwa	157	-92.4	-43.7
Lamwo	181	-95.0	-94.7
Luwero-1	157	-11.8	56.5
Luwero-2	334	-67.0	-66.8
Lwengo	143	-79.7	-78.6
Mayuge	439	-245.0	-237.7

Continued on next page

District	N	RBF	Mat-32
Moyo	578	-384.6	-350.9
Mubende-1	157	2.4	8.5
Mubende-2	289	-62.3	-55.7
Nebbi-1	339	-159.7	-147.1
Nebbi-2	145	-37.6	-37.2
Rubirizi	161	-153.2	-153.3
Rukungiri	544	-129.0	-116.4
Adjumani	528	-191.4	-179.6
Amolatar	335	-72.0	-53.7
Amuria	386	-168.7	-176.7
Bulambuli	171	-64.3	-62.5
Ibanda	305	-99.2	-94.6
Kalungu	151	-88.4	-71.7
Kiboga-1	338	-139.0	-138.9
Kiboga-2	160	-24.1	-23.4
Koboko	306	-136.7	-90.8
Kyenjojo-1	282	-26.7	-22.1
Kyenjojo-2	195	-140.1	-83.2
Manafwa	358	-79.1	-74.5
Masindi-1	186	-69.1	-40.6
Masindi-2	187	4.2	5.2
Masindi-3	130	-56.9	18.6
Napak	174	-140.6	-119.4
Ssembabule	578	-112.3	-108.8
Iganga-1	184	-49.2	-23.7
Iganga-2	177	46.0	46.1
Iganga-3	120	-15.1	-15.0
Kibaale	449	-133.2	-137.4
Mukono-1	348	369.0	373.0
Mukono-3	133	-28.2	-25.0
Buhweju	148	-43.0	-39.0
Buvuma	127	-72.3	-71.3
Kabale	571	-304.8	-303.7
Kapchorwa-1	153	-104.5	-64.1
Kapchorwa-2	233	-119.2	-60.6

Continued on next page

District	N	RBF	Mat-32
Kapchorwa-3	152	-130.1	-129.6
Kween	171	-142.2	-112.8
Lira-1	138	-75.0	-74.7
Lira-2	11	-1.0	-1.0
Lira-3	203	-71.9	-38.6
Lira-5	178	-181.8	-182.2
Mitooma	115	-128.5	-123.2
Nakasongola	487	-114.2	-111.2
Namutumba	295	-25.4	-32.1
Ntoroko	177	-130.1	-117.4
Otuke	186	-139.8	-117.4
Pallisa-1	182	-34.8	-33.0
Pallisa-2	174	-0.7	-0.5
Pallisa-3	129	-86.7	-89.6
Masaka-1	323	-13.7	-9.7
Masaka-4	128	-43.6	-40.8
Agago	176	-216.2	-159.9
Bududa	364	-17.4	16.8
Buikwe	125	0.5	-1.7
Bukedea	299	-214.2	-121.9
Kamuli-1	156	-48.0	-34.5
Kamuli-2	182	-23.0	-23.3
Kamuli-3	125	-58.9	-36.4
Kanungu	485	-153.5	-141.4
Kiryandongo	179	-35.5	-36.1
Kisoro	565	50.9	53.1
Kole	156	-67.5	-67.7
Mbale-1	166	53.7	71.8
Mbale-2	31	15.4	21.1
Mbale-3	376	-203.2	-203.2
Nakaseke	380	-98.8	-101.2
Ngora	176	-120.9	-84.2
Tororo-1	166	-58.4	-37.6
Tororo-2	395	-194.2	-194.1
Zombo	172	-40.1	-38.0

Continued on next page

District	N	RBF	Mat-32
Amuru	374	-127.0	-49.9
Bugiri-1	329	-79.8	-55.6
Bugiri-2	141	-131.1	-130.4
Bullisa	384	-195.5	-194.0
Kotido-1	162	-131.3	-110.4
Kotido-2	31	-6.0	-6.0
Kotido-3	270	23.7	24.8
Rakai-1	201	14.8	17.0
Rakai-2	358	-173.7	-102.6
Butambala	165	-70.7	-45.6
Kaabong	317	-55.8	-55.1
Kamwenge	559	-352.1	-290.7
Kiruhura	362	-225.3	-224.3
Kitgum-1	354	-18.1	-14.7
Kitgum-2	162	-108.7	-68.0
Kyankwanzi	135	-65.6	-66.7
Maracha	263	-102.8	-84.0
Nakapiripirit-1	365	-82.0	-80.3
Nakapiripirit-2	172	-68.4	-68.3
Sheema	187	-208.4	-161.3
Amudat	108	-73.1	-73.4
Budaka	293	-119.0	-117.3
Dokolo	390	-216.1	-148.1
Hoima	556	-192.2	-189.4
Kumi-1	192	85.3	88.5
Kumi-2	179	4.6	5.5
Kumi-3	110	-58.9	-58.0
Luuka	148	-71.0	-55.9
Mbarara-1	168	-61.4	-55.7
Mbarara-4	349	-30.8	-27.6
Moroto-1	363	-200.4	-126.1
Moroto-2	156	-94.8	-94.6
Oyam	358	-243.8	-196.5
Wakiso	561	-328.3	-328.8
Apac-1	178	-75.1	-59.9

Continued on next page

District	N	RBF	Mat-32
Apac-2	203	122.0	117.9
Apac-3	160	-89.6	-89.6
Bukomansimbi	145	-47.3	-46.7
Kibuku	182	-111.1	-102.3
Ntungamo	481	-147.2	-143.4

E.2.2 Addition of Linear Kernel

Table E.3 Linear kernel validation. The consecutive numbers at the end of the district names indicate a change in the districts delimitation. The second column shows the number of data points en each case. The third column indicates the kernel choosen to model the data, depending on the highest LOO-CV value. The fourth and fifth columns show the log LOO-CV scores for a base model (using a single kernel either Mat-32 or RBF) and a model that incorporates a linear kernel. When the number of observations available was less than 10, no comparison was done and a linear kernel was not used.

District	N	Kernel choosen	Base	Base+Linear
Abim	227	Mat-3/2 + linear	-100.4	-4.7
Adjumani	528	Mat-3/2 + linear	-179.6	-114.3
Agago	176	Mat-3/2 + linear	-159.9	-151.0
Alebtong	181	RBF + linear	-102.6	-83.4
Amolatar	335	Mat-3/2	-53.7	-199.3
Amudat	108	RBF + linear	-73.1	-71.5
Amuria	386	RBF	-168.7	-192.8
Amuru	374	Mat-3/2	-49.9	-111.8
Apac-1	178	Mat-3/2	-59.9	-68.0
Apac-2	203	RBF + linear	122.0	130.4
Apac-3	160	RBF + linear	-89.6	-53.1
Arua-1	162	RBF + linear	29.5	70.9
Arua-2	62	RBF + linear	33.1	35.6
Arua-3	307	Mat-3/2 + linear	-207.0	-148.2
Budaka	293	Mat-3/2 + linear	-117.3	-98.1
Bududa	364	Mat-3/2	16.8	-17.4
Bugiri-1	329	Mat-3/2	-55.6	-65.9

Continued on next page

District	N	Kernel choosen	Base	Base+Linear
Bugiri-2	141	Mat-3/2 + linear	-130.4	-87.1
Buhweju	148	Mat-3/2 + linear	-39.0	-9.2
Buikwe	125	RBF + linear	0.5	19.0
Bukedea	299	Mat-3/2 + linear	-121.9	-88.7
Bukomansimbi	145	Mat-3/2 + linear	-46.7	1.8
Bukwo	396	Mat-3/2 + linear	-220.1	-203.9
Bulambuli	171	Mat-3/2 + linear	-62.5	-28.1
Bullisa	384	Mat-3/2 + linear	-194.0	-172.8
Bundibugyo-1	381	Mat-3/2	-53.9	-107.4
Bundibugyo-2	190	Mat-3/2 + linear	-31.6	-11.2
Bushenyi-1	386	Mat-3/2 + linear	-54.4	-40.9
Bushenyi-4	24	Mat-3/2	8.4	2.9
Bushenyi-5	162	Mat-3/2 + linear	-75.3	-25.0
Busia	459	Mat-3/2 + linear	-85.5	-54.1
Butaleja	407	Mat-3/2	112.0	78.1
Butambala	165	Mat-3/2 + linear	-45.6	-31.1
Buvuma	127	Mat-3/2 + linear	-71.3	-70.7
Buyende	156	Mat-3/2	-62.0	-63.7
Dokolo	390	Mat-3/2	-148.1	-212.9
Gomba	148	Mat-3/2 + linear	-73.4	-61.6
Gulu-1	184	Mat-3/2 + linear	-89.2	-80.9
Gulu-2	198	Mat-3/2 + linear	4.7	11.3
Gulu-3	158	Mat-3/2 + linear	-16.5	27.0
Hoima	556	Mat-3/2	-189.4	-199.2
Ibanda	305	Mat-3/2 + linear	-94.6	-68.9
Iganga-1	184	Mat-3/2	-23.7	-37.6
Iganga-2	177	Mat-3/2	46.1	42.1
Iganga-3	120	Mat-3/2 + linear	-15.0	-7.9
Isingiro	307	Mat-3/2 + linear	-118.6	-90.3
Jinja	555	Mat-3/2 + linear	122.1	144.6
Kaabong	317	Mat-3/2	-55.1	-73.8
Kabale	571	Mat-3/2 + linear	-303.7	-200.0
Kabarole	562	Mat-3/2 + linear	-244.1	-122.9
Kaberamaido	512	Mat-3/2	-278.5	-366.8
Kalangala	433	RBF	-193.5	-204.6

Continued on next page

District	N	Kernel choosen	Base	Base+Linear
Kaliro	376	Mat-3/2	-206.8	-207.8
Kalungu	151	Mat-3/2	-71.7	-72.5
Kampala	565	Mat-3/2 + linear	-347.2	-248.7
Kamuli-1	156	Mat-3/2 + linear	-34.5	-12.9
Kamuli-2	182	RBF + linear	-23.0	58.0
Kamuli-3	125	Mat-3/2 + linear	-36.4	-19.6
Kamwenge	559	Mat-3/2 + linear	-290.7	-238.8
Kanungu	485	Mat-3/2 + linear	-141.4	-41.6
Kapchorwa-1	153	Mat-3/2	-64.1	-101.7
Kapchorwa-2	233	Mat-3/2	-60.6	-99.3
Kapchorwa-3	152	Mat-3/2 + linear	-129.6	-125.6
Kasese	489	Mat-3/2 + linear	-209.1	-123.7
Katakwi-1	145	Mat-3/2 + linear	-26.9	-2.5
Katakwi-2	295	Mat-3/2 + linear	-171.6	-164.6
Kayunga	565	Mat-3/2 + linear	50.0	124.7
Kibaale	449	RBF + linear	-133.2	-84.2
Kiboga-1	338	Mat-3/2 + linear	-138.9	-131.9
Kiboga-2	160	Mat-3/2	-23.4	-78.4
Kibuku	182	Mat-3/2 + linear	-102.3	-58.2
Kiruhura	362	Mat-3/2 + linear	-224.3	-198.9
Kiryandongo	179	RBF + linear	-35.5	36.4
Kisoro	565	Mat-3/2	53.1	46.5
Kitgum-1	354	Mat-3/2 + linear	-14.7	17.8
Kitgum-2	162	Mat-3/2 + linear	-68.0	-49.5
Koboko	306	Mat-3/2 + linear	-90.8	-81.0
Kole	156	RBF + linear	-67.5	-40.0
Kotido-1	162	Mat-3/2	-110.4	-124.3
Kotido-2	31	Mat-3/2 + linear	-6.0	-1.7
Kotido-3	270	Mat-3/2 + linear	24.8	91.2
Kumi-1	192	Mat-3/2	88.5	16.8
Kumi-2	179	Mat-3/2 + linear	5.5	10.3
Kumi-3	110	Mat-3/2	-58.0	-69.7
Kween	171	Mat-3/2 + linear	-112.8	-72.0
Kyankwanzi	135	RBF + linear	-65.6	-10.6
Kyegegwa	157	Mat-3/2	-43.7	-47.3

Continued on next page

District	N	Kernel choosen	Base	Base+Linear
Kyenjojo-1	282	Mat-3/2 + linear	-22.1	57.5
Kyenjojo-2	195	Mat-3/2 + linear	-83.2	-22.1
Lamwo	181	Mat-3/2 + linear	-94.7	-45.6
Lira-1	138	Mat-3/2 + linear	-74.7	-61.3
Lira-2	11	RBF	-1.0	-1.0
Lira-3	203	Mat-3/2	-38.6	-62.5
Lira-5	178	RBF + linear	-181.8	-136.4
Luuka	148	Mat-3/2 + linear	-55.9	-31.5
Luwero-1	157	Mat-3/2 + linear	56.5	94.1
Luwero-2	334	Mat-3/2 + linear	-66.8	-27.5
Lwengo	143	Mat-3/2	-78.6	-95.7
Lyantonde	248	Mat-3/2 + linear	-193.6	-160.9
Manafwa	358	Mat-3/2 + linear	-74.5	-40.9
Maracha	263	Mat-3/2 + linear	-84.0	-56.2
Masaka-1	323	Mat-3/2	-9.7	-10.6
Masaka-4	128	Mat-3/2 + linear	-40.8	0.5
Masindi-1	186	Mat-3/2	-40.6	-48.2
Masindi-2	187	Mat-3/2	5.2	-16.4
Masindi-3	130	Mat-3/2	18.6	-15.4
Mayuge	439	Mat-3/2 + linear	-237.7	-133.8
Mbale-1	166	Mat-3/2	71.8	37.0
Mbale-2	31	Mat-3/2	21.1	19.8
Mbale-3	376	RBF + linear	-203.2	-141.1
Mbarara-1	168	Mat-3/2	-55.7	-57.1
Mbarara-4	349	Mat-3/2 + linear	-27.6	68.1
Mitooma	115	Mat-3/2 + linear	-123.2	-106.5
Mityana	391	Mat-3/2 + linear	-69.8	-29.5
Moroto-1	363	Mat-3/2	-126.1	-203.8
Moroto-2	156	Mat-3/2 + linear	-94.6	-80.2
Moyo	578	Mat-3/2	-350.9	-356.7
Mpigi-1	349	Mat-3/2 + linear	-121.7	-84.9
Mpigi-3	163	RBF + linear	-15.4	9.1
Mubende-1	157	Mat-3/2 + linear	8.5	13.1
Mubende-2	289	Mat-3/2 + linear	-55.7	-2.8
Mukono-1	348	Mat-3/2	373.0	369.0

Continued on next page

District	N	Kernel choosen	Base	Base+Linear
Mukono-3	133	Mat-3/2 + linear	-25.0	4.4
Nakapiripirit-1	365	Mat-3/2 + linear	-80.3	-45.2
Nakapiripirit-2	172	Mat-3/2	-68.3	-115.4
Nakaseke	380	RBF + linear	-98.8	-80.2
Nakasongola	487	Mat-3/2 + linear	-111.2	-30.5
Namayingo	98	RBF + linear	-138.7	-78.4
Namutumba	295	RBF + linear	-25.4	44.0
Napak	174	Mat-3/2	-119.4	-140.9
Nebbi-1	339	Mat-3/2 + linear	-147.1	-142.4
Nebbi-2	145	Mat-3/2 + linear	-37.2	-26.0
Ngora	176	Mat-3/2 + linear	-84.2	-39.1
Ntoroko	177	Mat-3/2 + linear	-117.4	-106.7
Ntungamo	481	Mat-3/2	-143.4	-148.7
Nwoya	180	Mat-3/2 + linear	-101.8	-87.9
Otuke	186	Mat-3/2	-117.4	-143.4
Oyam	358	Mat-3/2 + linear	-196.5	-100.7
Pader-1	374	Mat-3/2	-109.7	-178.6
Pader-2	135	Mat-3/2 + linear	-87.4	-77.3
Pallisa-1	182	Mat-3/2 + linear	-33.0	-29.8
Pallisa-2	174	Mat-3/2	-0.5	-2.7
Pallisa-3	129	RBF + linear	-86.7	-75.4
Rakai-1	201	Mat-3/2 + linear	17.0	33.7
Rakai-2	358	Mat-3/2	-102.6	-138.0
Rubirizi	161	RBF + linear	-153.2	-67.9
Rukungiri	544	Mat-3/2	-116.4	-152.9
Serere	120	Mat-3/2	-37.0	-80.1
Sheema	187	Mat-3/2	-161.3	-233.4
Sironko-1	340	Mat-3/2 + linear	-38.7	-19.1
Sironko-2	149	Mat-3/2 + linear	-60.9	0.9
Soroti-1	361	Mat-3/2 + linear	-233.4	-226.5
Soroti-2	143	Mat-3/2	-52.9	-58.9
Ssembabule	578	Mat-3/2 + linear	-108.8	-18.8
Tororo-1	166	Mat-3/2	-37.6	-43.1
Tororo-2	395	Mat-3/2 + linear	-194.1	-80.1
Wakiso	561	RBF + linear	-328.3	-312.2

Continued on next page

District	N	Kernel choosen	Base	Base+Linear
Yumbe	499	Mat-3/2	-142.6	-143.5
Zombo	172	Mat-3/2 + linear	-38.0	-32.0

E.3 Outlier Analysis

Table E.4 Outliers detection. The consecutive numbers at the end of the district names indicate a change in the districts delimitation. The second column shows the number of observations available. The third column indicate the number of outliers defined after comparing the homoscedastic and heteroscedastic models. The last two columns show the log LOO-CV scores for the homoscedastic and heteroscedastic models, respectively. This outlier diagnostic was only applied when there were more than 50 observations available, otherwise an homoscedastic model was used.

District	Observations	Outliers	Homoscedastic	Heteroscedastic
Abim	227	7	17.96	26.96
Adjumani	528	14	-9.05	71.22
Agago	176	13	-52.89	-20.00
Alebtong	181	4	-53.56	-32.52
Amolatar	335	5	-15.31	-4.85
Amudat	108	5	-44.83	-6.25
Amuria	386	6	-98.71	-18.98
Amuru	374	6	0.87	28.81
Apac-1	178	4	-14.30	22.40
Apac-2	203	3	137.54	146.96
Apac-3	160	1	-50.58	-49.42
Arua-1	162	7	86.03	99.80
Arua-2	62	–	–	–
Arua-3	307	11	-26.43	98.72
Budaka	293	7	-10.97	27.84
Bududa	364	7	63.15	111.46
Bugiri-1	329	14	46.71	154.84
Bugiri-2	141	6	-56.02	-35.32
Buhweju	148	1	-1.25	-0.65
Buikwe	125	1	34.89	39.46

Continued on next page

District	Observations	Outliers	Homoscedastic	Heteroscedastic
Bukedea	299	5	-37.84	-5.68
Bukomansimbi	145	0	4.82	4.78
Bukwo	396	14	-90.34	55.74
Bulambuli	171	2	11.85	41.10
Bullisa	384	11	-55.49	57.53
Bundibugyo-1	381	4	-11.97	17.33
Bundibugyo-2	190	3	32.18	59.96
Bushenyi-1	386	9	84.13	163.80
Bushenyi-2	2	–	–	–
Bushenyi-3	1	–	–	–
Bushenyi-4	24	–	–	–
Bushenyi-5	162	5	20.93	51.24
Busia	459	12	54.67	162.34
Butaleja	407	8	143.80	190.50
Butambala	165	4	12.08	42.09
Buvuma	127	2	-50.30	-40.59
Buyende	156	2	-49.69	-44.96
Dokolo	390	8	-66.73	28.80
Gomba	148	2	-13.24	23.82
Gulu-1	184	4	-24.65	5.96
Gulu-2	198	3	63.60	106.60
Gulu-3	158	3	25.43	34.70
Hoima	556	9	-104.08	-67.06
Ibanda	305	6	-18.79	21.88
Iganga-1	184	2	16.31	40.71
Iganga-2	177	–	–	–
Iganga-3	120	2	10.73	16.38
Isingiro	307	5	-23.93	32.36
Jinja	555	16	304.56	396.92
Kaabong	317	11	3.26	79.60
Kabale	571	11	7.20	349.37
Kabarole	562	23	40.49	150.38
Kaberamaido	512	22	-132.45	6.47
Kalangala	433	9	-121.49	-96.31
Kaliro	376	8	-153.73	-133.71

Continued on next page

District	Observations	Outliers	Homoscedastic	Heteroscedastic
Kalungu	151	3	-48.05	-40.84
Kampala	361	12	144.62	225.91
Kamuli-1	156	2	21.15	23.60
Kamuli-2	182	4	68.84	79.55
Kamuli-3	125	4	21.41	22.35
Kamwenge	559	11	-45.49	71.21
Kanungu	485	11	74.95	181.43
Kapchorwa-1	153	2	-39.82	-31.06
Kapchorwa-2	233	7	-23.69	17.16
Kapchorwa-3	152	10	-79.88	-56.22
Kasese	489	10	-56.41	-18.39
Katakwi-1	145	3	48.04	70.36
Katakwi-2	295	12	-87.54	2.94
Kayunga	565	5	240.02	308.93
Kibaale	449	25	97.49	334.55
Kiboga-1	338	12	8.90	196.19
Kiboga-2	160	3	-9.21	-5.49
Kibuku	182	1	-21.20	-20.71
Kiruhura	362	10	-88.23	-46.78
Kiryandongo	179	3	63.30	84.99
Kisoro	565	4	136.93	190.27
Kitgum-1	354	5	48.76	54.08
Kitgum-2	162	5	2.61	29.17
Koboko	306	4	-1.94	78.18
Kole	156	–	–	–
Kotido-1	162	7	-86.86	-81.35
Kotido-2	31	–	–	–
Kotido-3	270	12	108.67	120.49
Kumi-1	192	2	123.94	153.33
Kumi-2	179	1	80.94	90.06
Kumi-3	110	3	-44.11	-37.78
Kween	171	3	-38.31	-19.97
Kyankwanzi	135	–	–	–
Kyegegwa	157	0	-42.04	-42.10
Kyenjojo-1	282	1	70.44	71.60

Continued on next page

District	Observations	Outliers	Homoscedastic	Heteroscedastic
Kyenjojo-2	195	2	-8.39	-2.43
Lamwo	181	2	-7.75	-2.98
Lira-1	138	2	-31.88	-13.52
Lira-2	11	–	–	–
Lira-3	203	6	8.75	43.78
Lira-4	1	–	–	–
Lira-5	178	13	-66.39	-0.70
Luuka	148	–	–	–
Luwero-1	157	1	98.48	100.22
Luwero-2	334	5	-2.63	10.07
Lwengo	143	4	-50.92	-38.91
Lyantonde	248	11	-73.18	41.39
Manafwa	358	6	11.77	38.40
Maracha	263	5	-47.51	-16.32
Masaka-1	323	17	90.67	188.62
Masaka-2	1	–	–	–
Masaka-3	5	–	–	–
Masaka-4	128	4	28.35	45.69
Masindi-1	186	5	11.06	61.93
Masindi-2	187	3	62.15	128.82
Masindi-3	130	1	41.00	56.64
Mayuge	439	12	-40.72	-9.63
Mbale-1	166	–	–	–
Mbale-2	31	–	–	–
Mbale-3	376	18	-6.73	210.78
Mbarara-1	168	3	-3.99	61.05
Mbarara-2	1	–	–	–
Mbarara-3	7	–	–	–
Mbarara-4	349	7	98.26	134.84
Mitooma	115	2	-88.58	-78.75
Mityana	391	6	20.08	39.53
Moroto-1	363	10	-70.83	-47.69
Moroto-2	156	8	-58.97	-52.65
Moyo	578	12	-135.62	136.19
Mpigi-1	349	9	-33.20	-17.02

Continued on next page

District	Observations	Outliers	Homoscedastic	Heteroscedastic
Mpigi-2	5	–	–	–
Mpigi-3	163	2	-18.38	-15.73
Mubende-1	157	2	30.51	34.36
Mubende-2	289	3	51.79	80.48
Mukono-1	348	7	421.98	594.88
Mukono-2	6	–	–	–
Mukono-3	133	1	52.00	62.33
Nakapiripirit-1	365	9	35.22	81.54
Nakapiripirit-2	172	5	-47.29	-41.01
Nakaseke	380	9	22.51	104.15
Nakasongola	487	9	86.26	157.64
Namayingo	98	3	-53.51	-18.23
Namutumba	295	2	48.92	50.16
Napak	174	5	-86.76	-57.20
Nebbi-1	339	17	-16.17	148.09
Nebbi-2	145	2	-14.60	-10.14
Ngora	176	2	-29.39	-26.02
Ntoroko	177	5	-58.45	-42.64
Ntungamo	481	9	-30.54	52.82
Nwoya	180	5	-14.51	22.20
Otuke	186	5	-62.29	-12.22
Oyam	358	10	-53.71	-2.99
Pader-1	374	5	-39.81	-3.70
Pader-2	135	–	–	–
Pallisa-1	182	2	29.28	40.77
Pallisa-2	174	2	4.66	4.84
Pallisa-3	129	3	-46.20	-14.39
Rakai-1	201	2	42.83	43.25
Rakai-2	358	7	-38.60	15.79
Rubirizi	161	2	-44.62	-18.71
Rukungiri	544	9	-1.17	101.77
Serere	120	1	-37.07	-37.07
Sheema	187	9	-94.69	10.80
Sironko-1	340	12	73.26	141.52

Continued on next page

District	Observations	Outliers	Homoscedastic	Heteroscedastic
Sironko-2	149	4	49.48	75.16
Soroti-1	361	5	-83.15	72.34
Soroti-2	143	2	-40.46	-36.55
Ssembabule	578	10	99.56	221.80
Tororo-1	166	8	18.56	78.68
Tororo-2	395	11	41.89	216.43
Wakiso	561	18	-85.22	145.17
Yumbe	499	18	6.79	169.09
Zombo	172	5	2.87	24.25