Gaussian processes for classification of spatial data in context of an early warning chain

Dipl.-Inform. Wirt Dominik Gallus

Karlsruhe Institute of Technology

A thesis submitted for the degree of

Doctor of Engineering (Dr.-Ing.)

2010 December

Karlsruhe, 26.10.2010

1. Reviewer: Prof. Dr.-Ing. Peter C. Lockemann

2. Reviewer: Prof. Dr. Mikhail Kanevski

Day of defense: 20.12.2010

Signature from head of PhD committee:

Zusammenfassung

Verarbeitung und Analyse von Daten mit Raum-/Zeitbezug mit dem Ziel einer Schätzung von Werten auf einer Menge von Datenpunkten, für welche keine Beobachtungen (Messungen) verfügbar sind ist Gegenstand mehrerer Teilgebiete der statistischen Wissenschaften. Dabei basiert die Abschätzung auf Stichproben, die aus einer Menge von Beispielen (Datenpunkten und Beobachtungen) bestehen. Das Spektrum der Anwendungen umfasst unterschiedliche Fragestellungen wie z.B. die Schätzung der Konzentration eines Minerals im Boden, die Schätzung der Verteilung von Schadstoffen in der Luft oder die Schätzung der Anfälligkeit gegenüber einer Naturgefahr und des damit verbundenen Risiko.

Gauss-Prozess-Techniken sind probabilistische Techniken, welche für Schätzung/ Vorhersage kontinuierlicher Werte verwendet werden. Der Grund hierfür liegt in der Handhabbarkeit mathematischer Ausdrücke im Fall kontinuierlicher Zielwerte. Im Gegensatz dazu ist die Anwendung von Gauss-Prozess-Techniken im Fall diskreter Zielwerte mit Mehraufwand verbunden, der durch Approximation hochdimensionaler Integrale über Produkte von Verteilungen unterschiedlichen Typs mit Hilfe deterministischer oder stochastischer Verfahren entsteht.

Ziel der Arbeit ist eine Untersuchung der Eignung von Gauss-Prozess-Techniken für Klassifikation (Schätzung diskreter Zielwerte) räumlicher Daten, mit Fokus auf Klassifikation der Gefährdung durch Massenbewegungen (Erdbewegungen, Schneelawinen). Dabei wird die Eignung von für die Schätzung/ Vorhersage räumlich verteilter Zielwerte bisher nicht angewandten Techniken am Beispiel hoch-dimensionaler realer Datensätze im Vergleich mit einer etablierten Technik des Maschinellen Lernens (Support Vector Machine (SVM)) überprüft , der gegenüber sie den Vorteil einer Aussage über die Unsicherheit in der Schätzung/ Vorhersage bieten, mit dem Potential, Entscheidungsunterstützung im Rahmen einer geeigneten Frühwarnkette zu verbessern.

Abstract

Processing and analysis of data describing the spatial distribution of quantities of interest aiming at estimation/ prediction of values at data points (locations) where observations (measurements) are missing has been topic of research in different fields of statistical science(s). Given a collection of data points with observations, quantities of interest may refer to the concentration of a particular mineral in a soil volume, concentration of pollutants within an area, incidence/ prevalence of a particular disease, or susceptibility to a particular kind of natural or hazard, and the corresponding risk.

Gaussian process techniques are probabilistic techniques commonly applied to prediction of continuous target values. This is due to analytical tractability of expressions involved in inference, with observations interpreted as an incomplete realization of a Gaussian process defined on the space of data points, transformed by a Gaussian noise process. In order to explain discrete target values, the assumption of a non-Gaussian process acting on the prior Gaussian process is introduced, resulting in intractable expressions. Consequently, classification problems have to be dealt with in a different (in general, more involving) way.

Aim of this work is an investigation of the applicability of Gaussian process classification techniques to prediction of categorical variables (classification) of spatial data on regional scale, focusing on occurence of mass movements (earth movements, snow avalanches). This is achieved by qualitative and quantitative evaluation, indicating predictive performance (sensitivity) comparable to the predictive performance (sensitivity) of the Support Vector Machine (SVM), with potential to improve decision support resulting from uncertainty estimates provided by Gaussian process techniques.

Declaration

This thesis describes work carried out between April 2007 and November 2010 at FZI Forschungszentrum Informatik.

I declare that this work was composed by myself and has not been submitted in any other application.

Acknowledgements

I would like to thank Prof. Peter C. Lockemann for the opportunity of an investigation into the topic of applicability of statistical/ probabilistic machine learning techniques (Gaussian process techniques) to spatial prediction (classification) problems. Without his support, this thesis would not have been possible.

I would like to thank Prof. Mikhail Kanevski (Université de Lausanne, Institut de géomatique et d'analyse du risque) for helpful discussions. His knowledge of topics in spatial prediction has proven invaluable in clarifying a range of questions.

Contents

Chapter 1

Introduction

Processing and analysis of data describing the spatial distribution of quantities of interest aiming at estimation/ prediction of values at data points (locations) where observations (measurements) are missing has been topic of research in different fields of statistical science(s). Given a collection of data points with observations, quantities of interest may refer to the concentration of a particular mineral in a soil volume, concentration of pollutants within an area, incidence/ prevalence of a particular disease, or susceptibility to a particular kind of natural hazard, and the corresponding risk.

Since the early work of Krige (20) and Matheron (24), geostatistics (4) has been established as a mainstream method for working with spatial data. Developed in the geological sciences for the task of estimation of concentration of mineral deposits (prediction of ore grade), the success of geostatistical techniques, based on recognition and modelling of spatial correlation, resulted in application to prediction problems in a range of domains, including the environmental sciences (meteorology, hydrology, ecology), epidemiology, geography, and a number of other fields.

In context of statistical prediction, recognition and modelling of correlation can be seen as a characteristic of geostatistical methods and a collection of different techniques developed in statistics (26) and machine learning (38), (35) to deal with problems involving spatial and non-spatial data. These techniques are capable of making use of information in a description of correlation between data points. In presence of correlation in data, data points convey information about each other, with explicit modelling of correlation between data points resulting in more accurate predictions.

Due to the focus on spatial location, geostatistics has focused on prediction problems where values of observations are assumed to be the outcomes of a (continuous) function of coordinates in low-dimensional (Euclidean) space (i.e., in $\mathbb{R}^n$, with $n = 2$, or $n = 3$). Hence, the design of traditional geostatistical procedures (involving estimation of correlation structure from data [1]) does not lend itself to more general spatial prediction problems, where values of observations (which need not be continuous) are assumed to depend on a set of D variables (*geo-features*), or to spatio-temporal problems. At this point, techniques developed in statistics (26) and machine learning (38), (35) introduce several advantages, including applicability to more complex prediction tasks, generalization to different/ more complex models (allowing for application to different prediction tasks, e.g. prediction of categorical (i.e., non-continuous) variables), more objective estimation of correlation parameters, and the possibility of introduction of techniques suitable to deal with larger data sets.

Aim of this work Aim of this work is an investigation of the applicability of statistical/ probabilistic machine learning techniques not previously applied in spatial prediction to the task of prediction of categorical variables (classification) of spatial data on regional scale. Specifically, a class of discriminative probabilistic techniques developed in statistics and machine learning, referred to as Gaussian process techniques, is investigated, focusing on occurence of mass movements (earth movements, snow avalanches). This problem is a particular instance of a classification problem, with values to be predicted representing class membership (i.e., whether a data point (location) is considered susceptible to a particular type of movement (in case of spatio-temporal problems, subject to mass movement hazard), or not). In context of hazard prediction, quantities of interest are defined to be probabilities of movement occurence, resulting in the special case of probabilistic classification. Due to the high-dimensional nature of the problem (with data points described by a set of D variables (with $D > 2$, in general)) and the type of values to be predicted, techniques developed in statistics and machine learning are considered, with focus on probabilistic techniques providing information related to uncertainty in predictions, of interest when pre-

[1] In geostatistics, this is referred to as the variography procedure.

dictions are made based on real-world data (where observations may be missing). The work summarizes results of research in context of project 'Development of suitable information systems for early warning systems' (EGIFF)[1], with focus on introduction of techniques aiming at improvements in processing of data in context of an early warning chain focusing on the occurence of movements.

In this work, these results consist of:

- Investigation of the applicability of approximate inference techniques not previously applied in spatial prediction to classification of high-dimensional spatial data on regional scale, focusing on classification of susceptibility to mass movements (earth movements) and prediction of avalanche hazard;

- Application of these techniques, with results indicating predictive performance comparable to established non-probabilistic techniques (Support Vector Machines), providing additional information related to uncertainty in prediction, with the potential to improve decision support;

- Implementation of these techniques in a way allowing for flexible use in context of a suitable early warning chain, extending to arbitrary classification tasks.

Outline of the thesis In the following, an outline of the thesis is given:

In chapter 2, spatial prediction is introduced, including the common problems of regression (prediction of continuous target values) and classification (prediction of discrete target values). Following a short introduction to the problem, investigation of techniques developed in statistics/ machine learning is motivated based on requirements of increasingly complex applications in spatial prediction. In this context, two types of techniques (probabilistic and non-probabilistic techniques) are discussed, focusing on a class of discriminative probabilistic techniques, referred to as Gaussian process techniques. In chapter 3, Gaussian process techniques for regression are introduced, starting with the definition of a stochastic process, common to methods developed in geostatistics, statistics, and machine learning. In the chapter, model-free kriging methods and model-based techniques

[1] part of the GEOTECHNOLOGIEN research program (http://www.geotechnologien.de), funded by the German Federal Ministry of Research and Education (BMBF)

developed in statistics and machine learning are introduced, with expressions for the Best Linear Unbiased Predictor (BLUP) and Best Predictor (BP) derived from the (probabilistic) model. In chapter 4, Gaussian process techniques for classification are described, focusing on problems resulting from the assumption of discrete target values. In the chapter, it is shown how these problems can be approached, resulting in approximate expressions for a predictive distribution, providing information related to uncertainty in prediction in addition to probabilistic predictions in case of discrete target values . In chapter 5, Gaussian process techniques for prediction in case of large data sets are described, focusing on algebraic techniques (reduced-rank approximations to the covariance matrix) and sparse Gaussian process techniques for classification. In chapter 6, the predictive performance (sensitivity) of Gaussian process techniques for classification is evaluated on two high-dimensional real-world spatial data sets, describing the occurence of different types of mass movements. Chapter 7 concludes, summarizing results of the work.

Chapter 2

Spatial prediction

Summary In this chapter, spatial prediction is introduced, including the common problems of regression (prediction of continuous target values) and classification (prediction of discrete target values) of spatial (spatio-temporal) data. Subsequently, investigation of techniques developed in statistics/ machine learning is motivated based on requirements of increasingly complex applications in spatial prediction, focusing on classification of spatial (spatio-temporal) data involving the occurence of hazardous mass movements (earth movements, snow avalanches). In this context, probabilistic and non-probabilistic techniques are considered, focusing on a class of discriminative probabilistic techniques (Gaussian process techniques), based on the argument of availability of a predictive distribution, providing information related to uncertainty in prediction, with potential to improve decision support.

The problem of spatial prediction is to build a model providing predictions for quantities of interest at any location in an area, given examples consisting of data points (denoted by $\mathbf{x}_i$, with $i = 1, \dots, N$), described by a set of variables and measurements/ observations t_i (target variables). Given examples $\{\mathbf{x}_i, t_i\}$ and a new data point $\mathbf{x}_{N+1}$, the goal of spatial prediction is to accurately estimate/ predict the value of target variable t_{N+1} at $\mathbf{x}_{N+1}$, based on information contained in $\{\mathbf{x}_i, t_i\}$.

In general, target variables can be continuous or non-continuous. Depending on the type of target variables, the spatial prediction problem is referred to as a regression problem (with $t_i \in \mathbb{R}$), ordinal (count type) regression problem (with

$t_i \in \mathrm{Z}^+$), or classification problem (with $t_i \in \{c_1, \ldots, c_k, \ldots, c_K\}$). Depending on the type of the problem, different techniques can be applied, with different techniques available for prediction problems of different types.

In context of spatial data, prediction of quantities of interest (values of targe variables) at data points (locations) has traditionally been performed within the framework of kriging techniques (4). However, application of kriging techniques has focused on data points described by coordinates in low-dimensional Euclidean space and continuous target values, i.e., on low-dimensional regression problems.

With recent technological advances, more data (sensoric measurements) has been made available, resulting in more complex prediction problems of different types. With more complex data sets (consisting of data points described by D variables (with $D > 2$, in general), and target variables of different types), resulting prediction problems suggest the application of more general prediction techniques developed in statistics (26) and machine learning (38), (35), with particular techniques (notably, the Support Vector Machine (SVM)) established as methods of choice for regression and classification in context of spatial data (18), as a result of good predictive performance on different prediction tasks.

The aim of this work is an investigation of the applicability of statistical/ probabilistic machine learning techniques not previously applied in spatial prediction to classification of spatial data, focusing on occurence of mass movements. In modelling the prediction problem, it is assumed that values of discrete target variables representing class membership (indicating whether movement occurence at a data point (location) has been registered) are dependent on N data points described by $D > 2$ variables (environmental factors), contributing to disposition (tendency) of mass to move, or triggering movement. In consequence, the prediction problem suggests the application of statistical/ machine learning techniques capable of modelling complex dependencies between the (categorical) target values and high-dimensional data points.

In the range of applicable methods developed in statistics and machine learning, available techniques can be divided in probabilistic and non-probabilistic techniques, corresponding to whether the assumption of a probabilistic model is made. This can be formalized as the assumption of (types for) probabilities contributing to the joint probability $p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \ldots, \mathbf{t}, t_{N+1}, \ldots)$ (with $\mathbf{t} = (t_1, \ldots, t_N)^T$, $t_i \in \{c_1, \ldots, c_K\}$), describing the assignment of target values to data points.

In general, non-probabilistic techniques, including Artificial Neural Networks (ANN) (2) and Support Vector Machines (SVM) (38) do not make the assumption of a probabilistic model. Instead, these techniques assume that examples in $\{\mathbf{x}_i, t_i\}$ are i.i.d. (independent, identically distributed) samples drawn from some distribution. Unfortunately, this means that expression(s) for probabilities contributing to the joint probability $p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \ldots, \mathbf{t}, t_{N+1}, \ldots)$ are not available. However, when modelling real-world data, it is preferable to know these expressions, which can be used to obtain a predictive distribution, describing the probability of the assignment of target values to a new data point $\mathbf{x}_{N+1}$, given the set of examples $\{\mathbf{x}_i, t_i\}$. In addition to probabilistic predictions based on the predictive distribution, uncertainty estimates provide additional information, which is of interest when real-world data is considered (where observations may be missing). In general, this makes probabilistic prediction techniques well-suited to probabilistic classification tasks if real-world spatial data is considered, with probabilistic predictions allowing for probabilistic mapping.

Given a probabilistic model (i.e., expression(s) for probabilities contributing to $p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \ldots, \mathbf{t}, t_{N+1}, \ldots)$ and a new data point $\mathbf{x}_{N+1}$, two approaches to probabilistic classification can be considered. In the first, referred to as the generative approach, $p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \mathbf{t}, t_{N+1} = c_k)$ (with $k = 1, \ldots, K$) can be written as a product of a class-conditional $p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1} | \mathbf{t}, t_{N+1} = c_k)$ and a prior density $p(\mathbf{t}, t_{N+1} = c_k)$ for observations $\mathbf{t}, t_{N+1}$. From class-conditional and prior density, a predictive distribution is obtained using Bayes' Theorem, resulting in

$$p(t_{N+1} = c_k | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \mathbf{t}) = \\ \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{x}_{N+1}|\mathbf{t},t_{N+1}=c_k)p(\mathbf{t},t_{N+1}=c_k)}{\sum_{k=1}^{K} p(\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{x}_{N+1}|\mathbf{t},t_{N+1}=c_k)p(\mathbf{t},t_{N+1}=c_k)} \frac{1}{p(\mathbf{t}|\mathbf{x}_1,\ldots,\mathbf{x}_N[,\mathbf{x}_{N+1}])}$$

The alternative approach, referred to as the discriminative approach, focuses on modelling the posterior distribution for observations given data points (the set of variables describing data points) $p(\mathbf{t}, t_{N+1} = c_k | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1})$ in the product $p(\mathbf{t}, t_{N+1} = c_k | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}) p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1})$ to obtain the predictive distribution $p(t_{N+1} = c_k | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \mathbf{t})$ more directly.

In this work, the discriminative approach is adopted. Specifically, a class of discriminative probabilistic techniques developed in statistics and machine learning, referred to as Gaussian process classification techniques, is investigated with respect to applicability to the task of classification of high-dimensional real-world spatial/ spatio-temporal data, focusing on the occurence of mass movements. Assuming that the vector of variables describing data points is high-dimensional, the approach circumvents the problem of density estimation for class-conditional densities $p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1} | \mathbf{t}, t_{N+1} = c_k)$, focusing on a Gaussian prior defined over a collection of latent (i.e., unobservable) random variables. For inference, the Gaussian prior is combined with a likelihood function, resulting in a posterior over the latent variables. Subsequently, different techniques can be applied to obtain a predictive distribution, allowing for probabilistic predictions.

In context of spatial prediction, Gaussian process techniques can be thought of in the framework of a generalization of geostatistical techniques to high-dimensional data sets, based on related assumptions. In analogy to geostatistical techniques, Gaussian process techniques allow to include information about correlation between data points (variables describing data points) in terms of a (symmetric, positive definite) covariance function c. Extending the capabilities of geostatistical techniques, Gaussian process techniques introduce several advantages, including applicability to more complex prediction tasks, generalization to different prediction problems (e.g., prediction of categorical (i.e., non-continuous) variables), more objective estimation of correlation parameters (in terms of parameters of the covariance function), and the possibility of introduction of techniques suitable to deal with larger data sets.

In the following chapters, Gaussian process techniques are introduced, starting with regression methods developed from geostatistical techniques (chapter 3). Subsequently, Gaussian process classification techniques developed in statistics and machine learning are introduced (chapter 4). In chapter 5, techniques suitable to deal with a large number of data points to include in prediction are considered.

Chapter 3

Gaussian process regression

Summary This chapter deals with Gaussian process techniques for regression (prediction of continuous target values). Starting with the definition of a stochastic process and the introduction of the covariance function, describing correlation between data points, Gaussian process-based techniques are introduced, with derivation of the Best Linear Unbiased Predictor (kriging predictor) from a measure of prediction error referred to as the mean square error (MSE). Subsequently, model-based techniques developed in statistics and machine learning are introduced, with expressions for the Best Linear Unbiased Predictor and Best Predictor derived from the model, making use of properties of the Gaussian. The chapter concludes with a method for estimation of parameters of the covariance function, based on optimization of the likelihood of data (observations) given the model. Throughout the chapter, it is shown how different approaches can be used to obtain optimal predictions within the framework of stochastic processes, with model-based techniques for regression linking geostatistical techniques to Gaussian process methods for classification, introduced in chapter 4.

3.1 Stochastic processes

All statistical techniques described in the following (chapter 3, 4, and 5) (including geostatistical, statistical, and machine learning/ Bayesian statistical techniques) are expressed in terms of stochastic processes, defined as follows:

Definition A stochastic process $Y(\cdot)$ is a collection of random variables $Y(\mathbf{x})$ $\{Y(\mathbf{x})|\mathbf{x} \in \mathbf{X}\}$ defined on an index set (input space) $\mathbf{X}$.

According to the Kolmogorov extension theorem (19), a stochastic process can be specified from a consistent (in terms of the marginalization property) collection of finite-dimensional probability distributions. Conversely, for a finite collection $\{Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)\}$, a probability distribution, referred to as the distribution of the process, can be obtained from the stochastic process:

$$p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)))$$
$$= \int \ldots \int p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N), Y(\mathbf{x}_{N+1}), \ldots, Y(\mathbf{x}_{N+n})))$$
$$dY(\mathbf{x}_{N+1}) \ldots dY(\mathbf{x}_{N+n}),$$

with $n \in \mathbb{N}$, and $(\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+n})^T \in \mathbf{X}^n$.

From above expression, $p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)))$ can be substituted for $Y(\cdot)$ if the finite collection $\{Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)\}$ is considered. Hence, if interested in the distribution of finite $\mathbf{y}$, it is possible to work with $(Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N))^T$ (and the corresponding probability distribution $p(\mathbf{y}) = p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)))$, not taking $(Y_{N+1}, \ldots, Y_{N+n})$ into account.

Within the stochastic process framework, examples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N, t_1, \ldots, t_N\}$ are interpreted as an incomplete realization of a stochastic process $Y(\cdot) := \{Y(\mathbf{x}) | \mathbf{x} \in \mathbf{X}\}$, transformed by a noise process $\{T_Y(y(\mathbf{x})) | y(\mathbf{x}) = Y(\cdot)\}$ acting on realizations of $Y(\cdot)$: [1]

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_N, t_1, \ldots, t_N\}$$
$$\sim p(T(\mathbf{x}_1), \ldots, T(\mathbf{x}_N) | Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)) p(Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N))$$

In general, the stochastic process underlying $\{\mathbf{t}, \mathbf{x}\} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N, t_1, \ldots, t_N\}$ is not assumed to be completely observable. Instead, $\{\mathbf{t}, \mathbf{x}\}$ is assumed to represent an (incomplete) realization of a transformed latent (i.e., unobservable) process $Y(\cdot)$.

Depending on the type of $T_Y(\cdot)$, different distributions of $\{\mathbf{t}, \mathbf{x}\}$ can be explained, resulting in prediction problems corresponding to regression problems (with $t_i \in \mathbb{R}$), count type regression problems (with $t_i \in \mathrm{Z}^+$), or classification problems (with $t_i \in \{c_1, \ldots, c_k, \ldots, c_K\}$ (with $k = 1, \ldots, K$)).

[1] see e.g. (7)

Given $\{\mathbf{t}, \mathbf{x}\}$, an optimal prediction for $T(\mathbf{x}_{N+1})$ at $\mathbf{x}_{N+1} \in \mathbf{X}$ can be made based on a predictive distribution, given by $p(T(\mathbf{x}_{N+1})|\mathbf{t})$ [1]. Depending on the type of $T_Y(\cdot)$, different techniques can be applied to obtain the predictive distribution, as described in this chapter, and in the sections in chapter 4.

3.1.1 The covariance function

In the stochastic process framework, correlation is formalized in terms of a symmetric, positive definite function $c(\mathbf{x}_i, \mathbf{x}_j)$ of input points $\mathbf{x}_i, \mathbf{x}_j$, referred to as the covariance function $cov(Y(\mathbf{x}_i), Y(\mathbf{x}_j))$ of the process, defined by

$c(\mathbf{x}_i, \mathbf{x}_j) = cov(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) = E((Y(\mathbf{x}_i) - E(Y(\mathbf{x}_i)))(Y(\mathbf{x}_j) - E(Y(\mathbf{x}_j)))),$

with $E(Y(\mathbf{x}))$ determined by the mean function $\mu(\mathbf{x})$.

A number of covariance functions is common in practice. E.g.,
the squared exponential covariance function c_{SE},

$c_{SE}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{r^2}{2l^2})$, with length scale $l > 0$, and (squared) input distance $r^2 = (\mathbf{x}_j - \mathbf{x}_i)^2$,

the Matern covariance function c_{Matern},

$c_{Matern} = \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{\sqrt{2\nu}r}{l})^{\nu} K_{\nu}(\frac{\sqrt{2\nu}r}{l})$

with smoothness parameter $\nu > 0$, length scale $l > 0$, $r = \|\mathbf{x}_j - \mathbf{x}_i\|$, and K_ν denoting the modified Bessel function [2],

the rational quadratic covariance function c_{RQ},

$c_{RQ} = (1 + \frac{r^2}{2\alpha l^2})^{-\alpha}$

with prior shape parameter $\alpha > 0$, length scale $l > 0$, $r^2 = (\mathbf{x}_j - \mathbf{x}_i)^2$,

[1] see e.g. (39)

[2] see (1), sec. 9.6

and the (polynomial) dot product covariance function $c_{polydot}$,

$c_{polydot} = (\mathbf{x}_i \mathbf{x}_j)^p$ with $p \in \mathrm{Z}^+$.

3.1.2 Properties of stochastic processes

Stationarity A stochastic process can be characterized by properties related to the behavior of the process with respect to (linear) transformations of $\mathbf{x}$ in the input space. If the distribution of the process is not changed if $\mathbf{x}$ is translated, the process is stationary. In this case, the covariance depends only on the difference $\mathbf{h} = \mathbf{x}_j - \mathbf{x}_i$:

$$cov(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) = c(\mathbf{x}_i, \mathbf{x}_j) = c(\mathbf{x}_j - \mathbf{x}_i)$$

The important case of second order stationarity occurs if for any two $\mathbf{x}_i$, $\mathbf{x}_j$, the distribution of the process depends only on $\mathbf{h}$, up to its second moment:

$$cov(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) = E(Y(\mathbf{x}_i)Y(\mathbf{x}_i + \mathbf{h})) - E(Y(\mathbf{x}_i))^2$$

Intrinsic stationarity occurs in the case when differences of observed values at two $\mathbf{x}_i$, $\mathbf{x}_j$ have mean 0 ($E((Y(\mathbf{x}_i) - Y(\mathbf{x}_i + \mathbf{h}))) = 0$) and the variance of the differences depends only on $\mathbf{h}$, for any two $\mathbf{x}_i, \mathbf{x}_j$. [1]

Isotropy A different property of the stochastic process is related to the behavior of the process with respect to rotations of input space. If the covariance function is a function of $\|\mathbf{h}\| = \|\mathbf{x}_j - \mathbf{x}_i\|$ only (irrespective of the direction), the process is said to be isotropic. Otherwise (i.e., if the distribution of data has a preferred direction), the process is anisotropic.

3.2 Elements of geostatistics

The property of (intrinsic) stationarity is a central property in geostatistical tradition. In geostatistics, the condition of intrinsic stationarity must be met to apply a measure of spatial correlation particular to geostatistical practice, referred to as the (semi-)variogram $\gamma(\mathbf{h})$:

[1] The assumption of intrinsic stationarity is referred to as the intrinsic *hypothesis* in geostatistics, see (24)

$\gamma(\mathbf{h}) = \frac{1}{2} var(Y(\mathbf{x}_i) - Y(\mathbf{x}_i + \mathbf{h}))$

Traditionally, geostatistics bases its predictions on the (semi-)variogram, not assuming the existence of a covariance function. If the covariance is assumed to exist, there is a close relationship between variogram and covariance:

$\gamma(\mathbf{h}) = \frac{1}{2} cov(Y(\mathbf{x}_i), Y(\mathbf{x}_i)) - cov(Y(\mathbf{x}_i), Y(\mathbf{x}_i + \mathbf{h})) + \frac{1}{2} cov(Y(\mathbf{x}_i + \mathbf{h}), Y(\mathbf{x}_i + \mathbf{h}))$

which reduces to $\gamma(\mathbf{h}) = cov(Y(\mathbf{x}_i), Y(\mathbf{x}_i)) - cov(Y(\mathbf{x}_i), Y(\mathbf{x}_i + \mathbf{h}))$ in case of second-order stationarity (in which $var(\mathbf{x})$ is constant).

Hence, if the covariance is assumed to exist, $\gamma(\mathbf{h})$ can be derived from above expression. Typically, this is not the case in geostatistical practice. Instead, the variogram is built based on a discrete approximation to $\gamma(\mathbf{h})$, known as the empirical variogram $\gamma_{emp}(\mathbf{h})$, in the variography procedure.

3.2.1 The kriging predictor

In geostatistics, the procedure of predicting the values of $T(\mathbf{x}_{N+1}), T(\mathbf{x}_{N+2}), \ldots$ at unobserved locations $\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \ldots$ based on an assumed variogram (or, if the covariance is assumed to exist, the covariance function), is known as kriging, after Daniel G. Krige, who, in the 1950s, introduced a technique making use of information contained in a description of correlation in order to improve prediction accuracy.

In classical statistics, the kriging predictor is referred to as a Best Linear Unbiased Predictor (BLUP) (17). It is linear since the predictor $\tilde{t}(\mathbf{x}_{N+1}) = \boldsymbol{\lambda}^T \mathbf{t}$ (with $\mathbf{t} = (T(\mathbf{x}_1), \ldots, T(\mathbf{x}_N))^T$) is a linear function of the vector of observations. With $E(\tilde{t}(\mathbf{x}_{N+1})) = E(T(\mathbf{x}_{N+1}))$, the kriging predictor is unbiased. Finally, it is best in the sense of having minimum mean squared error $MSE = E((\lambda^T \mathbf{t} - T(\mathbf{x}_{N+1}))^2)$ in the class of linear (unbiased) predictors.

There are different names for various types of kriging. In the following, the main types, known as ordinary kriging (kriging with unknown but constant mean) and universal kriging (kriging with trend model) are described.

Ordinary kriging In case of ordinary kriging, the mean $E(T(\mathbf{x}))$ is unknown. In order to obtain $\tilde{t}(\mathbf{x}_{N+1}) = \boldsymbol{\lambda}^T \mathbf{t}$, the MSE under the linear predictor $\tilde{t}(\mathbf{x}_{N+1}) = \boldsymbol{\lambda}^T \mathbf{t}$ is minimized, subject to

$$E(\tilde{t}(\mathbf{x}_{N+1})) = E(T(\mathbf{x}_{N+1})) \Leftrightarrow \textstyle\sum_{i=1}^N \lambda_i E(T(\mathbf{x}_i)) = \mu(\mathbf{x}_{N+1}) \Leftrightarrow \sum_{i=1}^N \lambda_i = 1$$
$$\Leftrightarrow \boldsymbol{\lambda}^T \mathbf{1} = 1$$

where $E(T(\mathbf{x}_i)) = \mu(\mathbf{x}) = \mu(\mathbf{x}_{N+1})$.

Universal kriging Universal kriging allows the mean of the distribution of the random process to be linear in a set of vectors $\mathbf{f}_i$ of values $f_1(\mathbf{x}_i), \ldots, f_P(\mathbf{x}_i)$, evaluated at locations $\mathbf{x}_i$. Often, the set of functions is chosen so that $\mathbf{F}\boldsymbol{\beta} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)^T \boldsymbol{\beta}$ describes a polynomial trend surface of order p (with $\mathbf{X} = \mathrm{I\!R}^D$, $D \in \{2, 3\}$, and $p << 10$, in general) for a $P \times 1$ vector $\boldsymbol{\beta}$:

$$\mathbf{f}_i^T \boldsymbol{\beta} = \beta_0 + \textstyle\sum_{j1=1}^D \beta_{j1} x_{j1} + \sum_{j1=1}^D \sum_{j2=1}^D \beta_{j1j2} x_{j1} x_{j2} + \ldots$$
$$+ \textstyle\sum_{j1=1}^D \ldots \sum_{jq=1}^D \beta_{j1\ldots jq} x_{j1} \ldots x_{jq}$$

In case of universal kriging, the MSE is minimized subject to

$$\textstyle\sum_{i=1}^N \sum_{j=1}^P \lambda_i f_j(\mathbf{x}_i) \beta_j = \sum_{j=1}^P f_j(\mathbf{x}_{N+1}) \beta_j \Leftrightarrow \boldsymbol{\lambda}^T \mathbf{F} \boldsymbol{\beta} = \mathbf{f}_{N+1}^T \boldsymbol{\beta} \forall \boldsymbol{\beta}$$

with the $P \times 1$ vector $\mathbf{f}_{N+1} = (f_1(\mathbf{x}_{N+1}), \ldots, f_P(\mathbf{x}_{N+1}))^T$, evaluated at data point $\mathbf{x}_{N+1}$.

3.2.1.1 Prediction

Prediction in terms of the variogram The expression $\boldsymbol{\lambda}_*$ for the kriging weights obtained by minimizing the MSE subject to the unbiasedness constraint (in terms of the variogram) for ordinary and universal kriging is derived in Appendix C.

Substituting $\boldsymbol{\lambda}_*$ in the linear predictor, the predicted value $\tilde{t}(\mathbf{x}_{N+1})$ at location $\mathbf{x}_{N+1}$ is

$$\boldsymbol{\lambda}_*^T \mathbf{t} \overset{(\mathbf{1}^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} \mathbf{1})^{-1} = \mathbf{A}}{=} (\boldsymbol{\gamma} + \mathbf{1}\mathbf{A}(1 - \mathbf{1}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}))^T \boldsymbol{\Gamma}^{-1} \mathbf{t}$$

in case of ordinary kriging, and $\boldsymbol{\lambda}_*^T \mathbf{t} \overset{(\mathbf{F}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{F})^{-1}=\mathbf{A}}{=} (\boldsymbol{\gamma} + \mathbf{F}\mathbf{A}(\mathbf{f}_{N+1} - \mathbf{F}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}))^T\boldsymbol{\Gamma}^{-1}\mathbf{t}$

in universal kriging,

where $\boldsymbol{\gamma} = (\gamma(\mathbf{x}_1 - \mathbf{x}_{N+1}), \dots, \gamma(\mathbf{x}_N - \mathbf{x}_{N+1}))^T \in \mathbb{R}^N$, and $\boldsymbol{\Gamma}$ is a $N \times N$ matrix with (i, j)-th element determined by $\gamma(\mathbf{x}_i - \mathbf{x}_j)$, for $i = 1, \dots, N$ and $j = 1, \dots, N$.

The achieved minimum MSE calculated using these weights (denoted MSE$_*$) is referred to as the kriging variance. Substituting $\boldsymbol{\lambda}_*$ into the expression for the MSE yields:

$$\mathrm{MSE}_* = 2\boldsymbol{\lambda}_*^T\boldsymbol{\gamma} - \boldsymbol{\lambda}_*^T\boldsymbol{\Gamma}\boldsymbol{\lambda}_* \overset{(\mathbf{1}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{1})^{-1}=\mathbf{A}}{=} \boldsymbol{\gamma}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - (1 - \mathbf{1}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})^T\mathbf{A}(1 - \mathbf{1}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})$$

in case of ordinary kriging, and $\mathrm{MSE}_* = 2\boldsymbol{\lambda}_*^T\boldsymbol{\gamma} - \boldsymbol{\lambda}_*^T\boldsymbol{\Gamma}\boldsymbol{\lambda}_* \overset{(\mathbf{F}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{F})^{-1}=\mathbf{A}}{=} \boldsymbol{\gamma}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - (\mathbf{f}_{N+1} - \mathbf{F}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})^T\mathbf{A}(\mathbf{f}_{N+1} - \mathbf{F}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})$

in universal kriging.

In geostatistical terms, the kriging variance is an estimate of the precision of prediction. Low kriging variance implies a high level of precision (high level of confidence) in prediction. Conversely, high kriging variance implies a low level of precision (low level of confidence) in prediction.

Prediction in terms of the covariance In geostatistical practice, the kriging procedure is typically performed in terms of the variogram. Assuming the covariance function is known, the MSE can be expressed in terms of the covariance:

$$MSE = var(T(\mathbf{x}_{N+1})) - 2\boldsymbol{\lambda}^T\mathbf{c} + \boldsymbol{\lambda}^T\mathbf{C}\boldsymbol{\lambda}$$

where $\mathbf{c} = (cov(T(\mathbf{x}_1), T(\mathbf{x}_{N+1})), \dots, cov(T(\mathbf{x}_N), T(\mathbf{x}_{N+1})))^T \in \mathbb{R}^N$, and $\mathbf{C}$ is a $N \times N$ matrix with entries c_{ij} determined by $cov(T(\mathbf{x}_i), T(\mathbf{x}_j))$, evaluated at locations $\mathbf{x}_i$, $\mathbf{x}_j$.

In this case, derivation of the kriging weights results in an alternative expression for the predictor $\tilde{t}(\mathbf{x}_{N+1})$ and MSE$_*$:

$$\boldsymbol{\lambda}_*^T\mathbf{t} \overset{(\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1})^{-1}=\mathbf{A}}{=} (\mathbf{c}+\mathbf{1A}(1-\mathbf{1}^T\mathbf{C}^{-1}\mathbf{c}))\mathbf{C}^{-1}\mathbf{t}$$

in case of ordinary kriging, and $\boldsymbol{\lambda}_*^T\mathbf{t} \overset{(\mathbf{F}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{F})^{-1}=\mathbf{A}}{=} (\mathbf{c}+\mathbf{FA}(\mathbf{f}_{N+1}-\mathbf{F}^T\mathbf{C}^{-1}\mathbf{c}))\mathbf{C}^{-1}\mathbf{t}$

in universal kriging.

3.3 Model-based statistics

Classical geostatistics (relying on variogram modelling and the use of kriging techniques) assumes no distributional model. A different approach to spatial statistics, borrowing ideas from classical statistics, makes predictions based on a particular distributional model. In contrast to the distribution-free technique, predictions under the distributional are based on a predictive distribution, describing the assignment of probabilities to predicted target values. Compared to the distribution-free technique, the assumption of a statistical model introduces several advantages, including the possibility of generalization to different/ more complex models (allowing for application to classification tasks, i.e., prediction of categorical variables), more objective estimation of correlation parameters (in terms of parameters of the covariance function) and the possibility of introduction of techniques suitable to deal with larger data sets.

In the following sections, two approaches to model-based spatial statistics are introduced. In the first, a model developed in classical statistics is presented. Subsequently, a fully non-parametric Bayesian approach to process-based inference, reflecting the point of view taken in machine learning, is described.

3.3.1 The linear model

In classical statistics, a model resulting in predictions equivalent to the kriging techniques is known as the (Gaussian) linear model. The linear model involving stochastic processes can be written

$$\mathbf{t} = \mathbf{F}\boldsymbol{\beta} + \mathbf{e}$$

with $\mathbf{t} = (T(\mathbf{x}_1), \ldots, T(\mathbf{x}_N))^T$, a $N \times P$ design matrix $\mathbf{F}$ of values of functions

$f_1, \ldots, f_P$ evaluated at data points (locations) $\mathbf{x}_i$ for $i = 1, \ldots, N$, a $P \times 1$ vector $\boldsymbol{\beta}$ of unknown fixed effects and a $N \times 1$ random vector $\mathbf{e}$, interpreted as an incomplete realization of a Gaussian process, with the zero function as mean, and covariance function c_e.

Due to the presence of the Gaussian process component, the linear model can be written

$$\mathbf{t} \sim N(\mathbf{t}|\mathbf{F}\boldsymbol{\beta}, \mathbf{C}_e)$$

with mean $\mathbf{F}\boldsymbol{\beta}$ and covariance matrix $\mathbf{C}_e$, with entries c_{ij} determined by the covariance function $c_e(\mathbf{x}_i, \mathbf{x}_j)$, evaluated at $\mathbf{x}_i$, $\mathbf{x}_j$ for $i = 1, \ldots, N$ and $j = 1, \ldots, N$.

Various forms of kriging can be accomodated in this framework. The models for simple and ordinary kriging are special cases with $P = 0$ and $P = 1$, respectively:

$$\mathbf{t} = \mathbf{e},$$

and

$$\mathbf{t} = \begin{pmatrix} f_1(\mathbf{x_1}) \\ f_1(\mathbf{x_2}) \\ \vdots \\ f_1(\mathbf{x_N}) \end{pmatrix} \beta + \mathbf{e} = f_1(\mathbf{x})\beta + \mathbf{e} \overset{f_1(\mathbf{x})\beta =: \boldsymbol{\mu}_T}{=} \boldsymbol{\mu}_T + \mathbf{e}$$

Simple kriging with zero mean, or (equivalently) constant mean $\boldsymbol{\mu}$ removed, can be expresed as $\mathbf{t} = \mathbf{e}$, or $\mathbf{t}_+ - \boldsymbol{\mu} = \mathbf{t} = \mathbf{e}$. Ordinary kriging (i.e., kriging with unknown but constant mean) can be expressed as $\mathbf{t} = \boldsymbol{\mu} + \mathbf{e}$.

With dimensionality D, universal kriging with a linear trend has $D + 1$ parameters. For a polynomial of order p in D dimensions, the number of parameters P grows proportionally to D^p with p ($< D^p$ due to symmetry)[1]. Hence, under the linear model, universal kriging with polynomial trend can be expressed as

[1] e.g., for $D = 2$ and $p = 2$, the surface takes the form $\beta_0 + \sum_{i=1}^{D} \beta_i x_i + \sum_{i=1}^{D} \sum_{j=1}^{D} \beta_{i,j} x_i x_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,1} x_1 x_1 + \beta_{1,2} x_1 x_2 + (\beta_{2,1} x_2 x_1) + \beta_{2,2} x_2 x_2$, with 3 additional (unique) terms due to the increase to $p = 2$.

$$\mathbf{t} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,D} & \dots & \dots & \dots & x_{1,1}^p & \dots & x_{1,D}^p \\ 1 & x_{2,1} & \dots & x_{2,D} & \dots & \dots & \dots & x_{2,1}^p & \dots & x_{2,D}^p \\ \vdots & \vdots & \dots & \vdots & & \dots & & \vdots & \dots & \vdots \\ 1 & x_{N,1} & \dots & x_{N,D} & \dots & \dots & \dots & x_{N,1}^p & \dots & x_{N,D}^p \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_P \end{pmatrix} + \mathbf{e}$$

, and, in the more general case of a set of functions $\{f_1(\mathbf{x}), \dots, f_P(\mathbf{x})\}$, as

$$\mathbf{t} = \begin{pmatrix} f_1(\mathbf{x}_1) & \dots & f_P(\mathbf{x}_1) \\ f_1(\mathbf{x}_2) & \dots & f_P(\mathbf{x}_2) \\ \vdots & \dots & \vdots \\ f_1(\mathbf{x}_N) & \dots & f_P(\mathbf{x}_N) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_P \end{pmatrix} + \mathbf{e}$$

3.3.1.1 Prediction under the linear model

Provided the form of the covariance function c_e is known, predictions under the linear model can be made in a way similar to kriging. In case when (additional) random effects are absent, prediction under the linear model is equivalent to (universal) kriging. However, in comparison to kriging, prediction under the linear model has the advantage that the covariance function(s) need not be stationary. In order for the variance of the predictor to be ≥ 0, it is sufficient for the covariance function of the process (the covariance matrix $\mathbf{C}_e$) to be positive definite:

$var(\sum_{i=1}^N \lambda_i T(\mathbf{x}_i)) \geq 0 \Leftrightarrow \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j cov(T(\mathbf{x}_i), T(\mathbf{x}_j)) \geq 0 \Leftrightarrow \boldsymbol{\lambda}^T \mathbf{C}_e \boldsymbol{\lambda} \geq 0 \forall \boldsymbol{\lambda} \in \mathrm{I\!R}^N$

In the following, two approaches resulting in predictors for the unknown value at a new location $\mathbf{x}_{N+1}$ are given. In the first, a Best Linear Unbiased Predictor under the linear model is derived. In the second approach, a Best Predictor (BP) is derived from the predictive distribution $p(T(\mathbf{x}_{N+1})|\mathbf{t})$, making use of properties of the Gaussian.

Best Unbiased Linear Predictor Derivation of the Best Linear Unbiased Predictor under the linear model is performed in a way similar to the (universal) kriging predictor (see Appendix C for details), by minimizing the MSE subject to an unbiasedness constraint.

Given a new data point $\mathbf{x}_{N+1}$ with known $\mathbf{f}_{N+1} = (f_1(\mathbf{x}_{N+1}), \dots, f_P(\mathbf{x}_{N+1}))^T$, the MSE can be written:

$$E((\boldsymbol{\lambda}^T\mathbf{t} - T(\mathbf{x}_{N+1}))^2) \overset{\boldsymbol{\lambda}^T\mathbf{F}\boldsymbol{\beta}=\mathbf{f}_{N+1}\boldsymbol{\beta}}{=} E(((\boldsymbol{\lambda}^T\mathbf{t} - \boldsymbol{\lambda}^T\mathbf{F}\boldsymbol{\beta}) - (T(\mathbf{x}_{N+1}) - \mathbf{f}_{N+1}\boldsymbol{\beta}))^2)$$
$$= \boldsymbol{\lambda}^T\mathbf{C}_e\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\mathbf{c} + var(T(\mathbf{x}_{N+1})),$$

with $\mathbf{c} = cov(\mathbf{t}, T(\mathbf{x}_{N+1}))$.

Introducing a $P \times 1$ vector $\boldsymbol{\alpha}$ of Lagrange multipliers α_k, $k = 1, \ldots, P$, the expression to be minimized subject to the unbiasedness constraint

$E(\tilde{t}(\mathbf{x}_{N+1})) = \boldsymbol{\lambda}^T E(\mathbf{t}) \Leftrightarrow \boldsymbol{\lambda}^T\mathbf{F} = \mathbf{f}_{N+1}^T$ is

$$\boldsymbol{\lambda}^T\mathbf{C}_e\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\mathbf{c} + var(T(\mathbf{x}_{N+1})) - 2\boldsymbol{\alpha}^T(\mathbf{F}^T\boldsymbol{\lambda} - \mathbf{f}_{N+1})$$

Differentiating the expression with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ and equating to 0 results in the matrix form

$$\begin{pmatrix} \mathbf{C}_e & -\mathbf{F} \\ \mathbf{F}^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ -\mathbf{f}_{N+1} \end{pmatrix}$$

Assuming that $\mathbf{C}_e$ is non-singular (this is the case when the covariance function is positive definite), and using the result for the inverse of a partitioned matrix $\mathbf{M}$ (see Appendix B for details), the solution can be written

$$\begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} \mathbf{P} & -\mathbf{C}_e^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})^{-1} \\ -(\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}_e^{-1} & -(\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ -\mathbf{f}_{N+1} \end{pmatrix}$$

with $\mathbf{P} = \mathbf{C}_e^{-1} - \mathbf{C}_e^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}_e^{-1}$.

Hence,

$$\boldsymbol{\lambda}_* = \mathbf{C}_e^{-1}(\mathbf{c} + \mathbf{F}(\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})^{-1}(\mathbf{f}_{N+1} - \mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{c}))$$

which results in the expression for the Best Linear Unbiased Predictor in the case of universal kriging, if the covariance function is used instead of the variogram function:

$$\boldsymbol{\lambda}_*^T\mathbf{t} \overset{(\mathbf{F}^{\mathrm{T}}\mathbf{C}_{\mathrm{e}}^{-1}\mathbf{F})^{-1}=\mathbf{A}}{=} (\mathbf{c} + \mathbf{F}\mathbf{A}(\mathbf{f}_{N+1} - \mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{c}^T))\mathbf{C}_e^{-1}\mathbf{t}$$

As in the case of kriging, $\boldsymbol{\lambda}_*$ can be substituted into the expression for the MSE, resulting in

$$\text{MSE}_* \overset{(\mathbf{F}^{\mathrm{T}}\mathbf{C}_e^{-1}\mathbf{F})^{-1}=\mathbf{A}}{=} var(T(\mathbf{x}_{N+1})) - \mathbf{c}^T\mathbf{C}_e^{-1}\mathbf{c} + (\mathbf{f}_{N+1} - \mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{c}^T\mathbf{A}(\mathbf{f}_{N+1} - \mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{c})$$

which is equivalent to the expression for the minimum MSE in case of universal kriging, if the covariance function is used instead of the variogram function.

Best Predictor A mathematically less involving way to obtain an expression equivalent to the BLUP results from the assumption of Gaussianity. Making use of the result for the conditional Gaussian distribution $p(\mathbf{x}_1|\mathbf{x}_2)$, given the joint Gaussian distribution $p(\mathbf{x}_1, \mathbf{x}_2)$ (see Appendix A for details), given by

$$\begin{pmatrix} T(\mathbf{x}_{N+1}) \\ \mathbf{t} \end{pmatrix} \sim N(\begin{pmatrix} T(\mathbf{x}_{N+1}) \\ \mathbf{t} \end{pmatrix} | \begin{pmatrix} \mathbf{f}_{N+1}^T\boldsymbol{\beta} \\ \mathbf{F}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} c & \mathbf{c} \\ \mathbf{c}^T & \mathbf{C}_e \end{pmatrix})$$

with $c = var(T(\mathbf{x}_{N+1}))$, and $\mathbf{c} = cov(T(\mathbf{x}_{N+1}), \mathbf{t})$, the predictive distribution $p(T(\mathbf{x}_{N+1})|\mathbf{t})$ is obtained:

$$p(T(\mathbf{x}_{N+1})|\mathbf{t}[, \boldsymbol{\beta}]) = N(T(\mathbf{x}_{N+1})|\mathbf{f}_{N+1}^T\boldsymbol{\beta} + \mathbf{c}\mathbf{C}_e^{-1}(\mathbf{t} - \mathbf{F}\boldsymbol{\beta}),$$
$$var(T(\mathbf{x}_{N+1})) - \mathbf{c}\mathbf{C}_e^{-1}\mathbf{c}^T)$$

A result equivalent to BLUP and MSE$_*$ is obtained by recognizing the dependence on $\boldsymbol{\beta}$ in $p(T(\mathbf{x}_{N+1})|\mathbf{t}[, \boldsymbol{\beta}])$. Emphasizing the dependence and making use of the result for the marginal distribution $p(\mathbf{y})$ given a marginal distribution $p(\mathbf{x})$ and a conditional distribution $p(\mathbf{y}|\mathbf{x})$ (with $p(\mathbf{x}) = p(\boldsymbol{\beta}|[\mathbf{t}]) \sim N(\boldsymbol{\beta}|E(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}, E((\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T) = (\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})^{-1})$ [1] , and $p(\mathbf{y}|\mathbf{x}) = p(T(\mathbf{x}_{N+1})|\mathbf{t}, \boldsymbol{\beta})$; see Appendix A for details), the predictive distribution is

$$p(T(\mathbf{x}_{N+1})|\mathbf{t})$$
$$= N(T(\mathbf{x}_{N+1})|(\mathbf{f}_{N+1}^T - \mathbf{c}\mathbf{C}_e^{-1}\mathbf{F})\hat{\boldsymbol{\beta}} + \mathbf{c}\mathbf{C}_e^{-1}\mathbf{t},$$
$$var(T(\mathbf{x}_{N+1})) - \mathbf{c}\mathbf{C}_e^{-1}\mathbf{c}^T + (\mathbf{f}_{N+1}^T - \mathbf{c}\mathbf{C}_e^{-1}\mathbf{F})(\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})^{-1}(\mathbf{f}_{N+1}^T - \mathbf{c}\mathbf{C}_e^{-1}\mathbf{F})^T)$$

i.e., a Gaussian with BLUP and MSE$_*$ as first and second moment, respectively.

[1] with $\hat{\boldsymbol{\beta}}$ denoting the BLUE/ GLS estimate $\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_{ML} = (\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{V}^{-1}\mathbf{t}$ obtained from maximizing $p(\mathbf{t}|\boldsymbol{\beta}) = (2\pi)^{-\frac{N}{2}}|\mathbf{C}_e|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\mathbf{t} - \mathbf{F}\boldsymbol{\beta})^T\mathbf{C}_e^{-1}(\mathbf{t} - \mathbf{F}\boldsymbol{\beta}))$ with respect to $\boldsymbol{\beta}$.

3.3.2 The Gaussian process model (GPM)

Application of the linear model for the task of predicting the value of $T(\mathbf{x}_{N+1})$ provides an alternative to the kriging procedure. A different approach to the problem is taken in the Bayesian framework.

In the (fully non-parametric) Bayesian approach, a prior probability distribution is placed directly over $\mathbf{y}$, interpreted as an incomplete realization of a stochastic process $Y(\cdot)$. In general, it is assumed that $Y(\cdot)$ is a latent Gaussian process with zero mean and covariance function reflecting assumptions about characteristics (continuity, differentiability) of functions (realizations of $Y(\cdot)$) involved in inference:

$$p(\mathbf{y}) = N(\mathbf{y}|\mathbf{0}, \mathbf{C})$$

with $\mathbf{y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N))^T$.

To explain the distibution of $\mathbf{t} = (T(\mathbf{x}_1), \ldots, T(\mathbf{x}_N))^T$, the prior is combined with a likelihood function $p(\mathbf{t}|\mathbf{y})$, reflecting assumptions about the type of a noise process $T_Y(\cdot)$. In Gaussian process regression, typically isotropic Gaussian likelihood is assumed (equivalent to the assumption of i.i.d Gaussian noise):

$p(\mathbf{t}|\mathbf{y}) = N(\mathbf{t}|\mathbf{y}, \sigma^2\mathbf{I}_N)$, with $\sigma^2\mathbf{I}_N = diag(\sigma^2)$[1], $\sigma^2 > 0$.

Making use of the result for the marginal distribution $p(\mathbf{y})$ given a marginal Gaussian distribution $p(\mathbf{x})$ and a conditional Gaussian distribution $p(\mathbf{y}|\mathbf{x})$ (see Appendix A for details), the Gaussian marginal can be written

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = N(\mathbf{0}, \mathbf{K} = \mathbf{C} + \sigma^2\mathbf{I}_N)$$

Based on the Gaussian marginal $p(\mathbf{t})$, inference in the Gaussian process model is performed by making use of the result for the conditional Gaussian distribution $p(\mathbf{x}_1|\mathbf{x}_2)$, given the joint Gaussian distribution $p(\mathbf{x}_1, \mathbf{x}_2)$:

$$\begin{pmatrix} T(\mathbf{x}_{N+1}) \\ \mathbf{t} \end{pmatrix} \sim N(\begin{pmatrix} T(\mathbf{x}_{N+1}) \\ \mathbf{t} \end{pmatrix} | \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} k & \mathbf{k} \\ \mathbf{k}^T & \mathbf{K} \end{pmatrix})$$

[1] with $A = diag(a)$ denoting a diagonal matrix with entries a.

with $k = var(T(\mathbf{x}_{N+1}))$, and $\mathbf{k} = cov(T(\mathbf{x}_{N+1}), \mathbf{t})$, the predictive distribution is obtained:

$$p(T(\mathbf{x}_{N+1})|\mathbf{t}) = N(T(\mathbf{x}_{N+1})|\mathbf{k}(\mathbf{C} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{t}),$$
$$var(T(\mathbf{x}_{N+1})) - \mathbf{k}(\mathbf{C} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{k}^T)$$

Prediction under the GPM is equivalent to prediction resulting from an application of simple kriging (assuming i.i.d. Gaussian noise). In contrast to the linear model, the assumption of a deterministic trend $\mathbf{F}\boldsymbol{\beta}$ in the mean is dropped , reducing the increased computational complexity (in the order of P^3), resulting from inversion of the $P \times P$ matrix product $(\mathbf{F}^T\mathbf{C}_e^{-1}\mathbf{F})$.

3.3.3 Hyperparameter estimation

Application of the model-based statistical techniques to spatial data provides an alternative to kriging techniques, resulting in predictive distributions for quantities of interest. Another advantage of model-based statistical techniques in comparison to kriging is a more principled/ objective method of estimating the parameters of the correlation structure, modelled by the covariance function. In the linear model and the Gaussian process model, this is achieved by maximizing the likelihood of the data (observations $\mathbf{t}$) with respect to the vector of parameters $\boldsymbol{\theta}$ of the covariance function. In general, the procedure results in an optimal estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. From the statistical point of view, the technique has the properties of statistical consistency and asymptotic normality [1].

Automatic Relevance Determination The technique of hyperparameter estimation can be extended by including a separate parameter for each input variable $\mathbf{x}_j$ in the covariance function. Then, optimization of parameters allows the relative importance of covariance parameters to be estimated from the data. I.e., if the covariance function is parametrized with a D-dimensional vector $\boldsymbol{\theta}$ (with component θ_j for each input variable $\mathbf{x}_j$), the procedure allows to determine the relevance of each $\mathbf{x}_j$ with respect to the target variables.

[1] i.e., $\lim_{N\to\infty} P(|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}| \geq \epsilon) = 0$ for $\epsilon > 0$, and, for $N \to \infty$, $p(\hat{\boldsymbol{\theta}}_N) \to N(\hat{\boldsymbol{\theta}}_N|\boldsymbol{\theta}, \mathcal{I}(\hat{\boldsymbol{\theta}}_N))$, with $\hat{\boldsymbol{\theta}}_N$ denoting the ML estimator based on a sample of size N, and $\mathcal{I}(\hat{\boldsymbol{\theta}}_N)$ denoting the Fisher information matrix, evaluated at $\hat{\boldsymbol{\theta}}_N$.

maximum likelihood In the maximum likelihood approach, the log likelihood of the data given the model $\boldsymbol{\theta}$ is given by

$$l_{ML} = \log p(\mathbf{t}) = \log((2\pi)^{-\frac{N}{2}}|(\mathbf{C} + \sigma^2\mathbf{I}_N)|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\mathbf{t}^T(\mathbf{C} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{t})$$
$$= -\tfrac{N}{2}\log(2\pi) - \tfrac{1}{2}\log(|(\mathbf{C} + \sigma^2\mathbf{I}_N)|) - \tfrac{1}{2}\mathbf{t}^T(\mathbf{C} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{t}$$

Making use of the result for the derivative of the inverse $\mathbf{C}_e^{-1}$ and the derivative of the log of the determinant $\log(|\mathbf{C}_e|)$ (see Appendix B.3 for details), the vector of estimates $\hat{\boldsymbol{\theta}}$ is obtained by differentiating the log likelihood function with respect to each hyperparameter θ_j, resulting in

$$\frac{\partial l_{ML}}{\partial \theta_j} = \frac{\partial \log p(\mathbf{t})}{\partial \theta_j} = -\tfrac{1}{2}tr(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \theta_j}) + \tfrac{1}{2}\mathbf{t}^T\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \theta_j}\mathbf{C}^{-1}\mathbf{t}$$

Having obtained an expression for the log likelihood and the partial derivatives for each θ_j, maximization of the log likelihood can be performed using efficient gradient-based optimization algorithms, e.g. conjugate gradient, or Quasi-Newton methods (29).

Chapter 4

Gaussian process classification

Summary In this chapter, Gaussian process techniques for classification are introduced, in preparation for application in chapter 6. In contrast to techniques for regression introduced in chapter 3, the distribution of (discrete) target values in classification implies the choice of a different probabilistic model, resulting in intractable expressions for distributions involved in inference. In this chapter, it is shown how these expressions can be approximated by means of deterministic or stochastic techniques developed in statistics and machine learning, resulting in approximate expressions for the predictive distribution, providing information related to uncertainty in addition to probabilistic predictions in case of discrete target values.

The techniques presented in chapter 3 are regression techniques. These techniques are suitable to deal with problems where observations t_i and the unknown value $T(\mathbf{x}_{N+1})$ to be predicted are continuous (with $\mathbf{t} \in \mathbb{R}^N$, $T(\mathbf{x}_{N+1}) \in \mathbb{R}$) and the vector $(\mathbf{t}, T(\mathbf{x}_{N+1}))$ can be assumed to follow the distribution of a Gaussian process. A different class of problems is classification, where, for all $\mathbf{x} \in \mathbf{X}$, $T(\mathbf{x})$ $\in \{c_1, \ldots, c_K\}$ and $T(\mathbf{x}_{N+1})$ is to be assigned an element of $\{c_1, \ldots, c_K\}$. From this definition, $T(\cdot)$ cannot be modelled by the Gaussian distribution alone. Consequently, classification problems have to be dealt with in a different (in general, more involving) way.

Within the stochastic process framework, probabilistic classification techniques have been developed in classical statistics and machine learning/ Bayesian statistics. In geostatistics, a technique referred to as indicator kriging has been applied to the task of modelling the (spatial) cumulative distribution of quantities of interest. In classical statistics, work on extending the linear model has led to development of the generalized linear mixed model (GLMM) (26), a generalization applicable to classification tasks in context of correlated data. In Bayesian statistics, stochastic approximations using Markov Chain Monte Carlo (MCMC) techniques (13), (22) are applied when distributions involved in inference are non-conjugate [1] [2] Also within the Bayesian framework, variational approximation methods have been introduced in machine learning, improving on techniques developed in classical statistics.

This chapter is structured as follows: In the next section, the geostatistical approach to probabilistic classification is introduced. Subsequently, classical and Bayesian model-based approaches are reviewed. In the first part, a GLMM suitable for prediction within thea stochastic process framework is developed by introducing the generalized linear model (GLM)(25), an extension of the linear model which allows the outcome of the model to result from a non-linear transformation of the linear predictor $\mathbf{F}\boldsymbol{\beta}$. Subsequently, the GLM is extended by appending a vector of random effects to the linear predictor, resulting in a general prediction technique capable of making use of (a description of) correlation. In the second part, the Bayesian approach to Gaussian process classification is described. As in the case of Gaussian process regression, a fully non-parametric point of view is taken, with inference taking place in the space of functions (realizations of a stochastic process). However, in contrast to the regression case, the likelihood function (reflecting the effect of a noise process $T_Y(\cdot)$) is not assumed to be Gaussian. In general, this precludes analytical treatment of expressions (integrals) involved in inference and approximate inference must be performed.

Approximate inference Both in the classical and the Bayesian statistical approach to Gaussian process classification, the generalization to a non-Gaussian likelihood introduces computational problems. Since the likelihood is no longer Gaussian, integrals involving realizations of a Gaussian process prior and the likelihood can no longer be evaluated in closed form. Hence, application of Gaussian processes to classification requires additional treatment, involving analytical, numerical, or stochastic approximations.

[1] In Bayesian statistics, a prior distribution $p(\boldsymbol{\theta})$ is said to be conjugate to a likelihood function $p(\mathbf{x}|\boldsymbol{\theta})$ if the resulting posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ is of the same type as $p(\boldsymbol{\theta})$.

[2] In particular, the Gaussian distribution is self-conjugate.

4.1 Geostatistical classification

The geostatistical techniques introduced in chapter 3 are regression/ interpolation techniques. For classification tasks, a variant of kriging referred to as indicator kriging has been used for classification (more generally, to estimate the probability of a value exceeding a certain threshold). The technique is based on a transformation of the data so that for a fixed threshold z_k, the values of the random variables $T(\mathbf{x}_i)$ (with $1 \leq i \leq N$) are replaced by values of indicator variables $I(\mathbf{x}_i)$, with

$$I(\mathbf{x}_i) = \begin{cases} 1 & \text{if } T(\mathbf{x}_i) > z_k \\ 0 & \text{otherwise} \end{cases}$$

After transformation, the variography procedure is performed as in the case of regression, resulting in a variogram model.

Based on the variogram model, ordinary kriging is performed on the values of the indicator variables, with predictions at locations $\mathbf{x}$ interpreted as probabilities $p(I(\mathbf{x}) = 1)$ (more generally, $p(T(\mathbf{x}) > z_k)$). The procedure can be iterated for a sequence of thresholds z_k for $k = 1, \ldots, K$, with $t_1 < \ldots < t_K$ and the result can be interpreted as an approximation to the (complementary) cumulative distribution function of the indicator variable at location $\mathbf{x}$.

In essence, indicator kriging is ordinary kriging (i.e., a regression technique) applied to binary class labels $I(\mathbf{x}) \in \{0, 1\}$. In particular, the technique lacks a probabilistic model reflecting the assumption of a (non-Gaussian) process $T_Y(\cdot)$, resulting in a mapping $\mathbb{R}^N \rightarrow \{c_1, \ldots, c_K\}^N$ for $\mathbf{y}$. Consequently, it is more appropriately referred to as a label regression technique. From a practical point of view, the technique suffers from the same drawbacks as ordinary kriging, due to limitations resulting from the use of the variogram (more generally, the variography procedure).

4.2 Model-based classification

4.2.1 The generalized linear model

In the linear model introduced in the previous chapter, the vector of observations is interpreted as an incomplete realization of a Gaussian process:

$\mathbf{t} \sim N(\mathbf{t}|\mathbf{F}\boldsymbol{\beta}, \mathbf{C}_e)$

In the generalized linear model, $\mathbf{t}$ is assumed to follow a distribution from an exponential family:

$p(\mathbf{t}) = \prod_{i=1}^{N} p(t_i) = \prod_{i=1}^{N} \exp(\frac{t_i\gamma_i - b(\gamma_i)}{\tau} - c(t_i, \tau))$

A second assumption introduced in the GLM is the assumption that the mean $E(\mathbf{t}) = \boldsymbol{\mu}$ of the distribution $p(\mathbf{t})$ need not equal the linear predictor $\boldsymbol{\eta} = \mathbf{F}\boldsymbol{\beta}$. Instead, the less restrictive structural assumption $E(\mathbf{t}) = \boldsymbol{\mu} = h(\boldsymbol{\eta}) = h(\mathbf{F}\boldsymbol{\beta})$, or, equivalently, $\mathbf{F}\boldsymbol{\beta} = \boldsymbol{\eta} = g(\boldsymbol{\mu})$ is made, with a suitable non-linear response function h, or, equivalently, a suitable link function $g = h^{-1}$.

Using h, the linear predictor can be transformed in a different range, depending on the task. Different response functions (corresponding to different exponential families for $p(t_i)$) can be used when modelling probabilities, counts, or continuous target variables. This way, the GLM can be used for regression (with $t_i \in \mathbb{R}$), probabilistic classification (with $p(t_i = c_k) \in [0, 1]$ for $k = 1, \ldots, K$), or other prediction tasks.

In (binary) classification, t_i is assumed to follow the Bernoulli distribution (denoted $Bin(1, \pi_i)$), modelling the probability of an outcome k ($k \in \{0, 1\}$) in a binary trial. In canonical form, the distribution can be written

$p(t_i = k) = \binom{1}{k}\pi_i^{t_i}(1 - \pi_i)^{1-t_i}$
$= \exp(\frac{t_i \log(\frac{\pi_i}{1-\pi_i}) - \log(1+\exp(\log(\frac{\pi_i}{1-\pi_i})))}{1})$,

with $\gamma_i = \log(\frac{\pi_i}{1-\pi_i})$, $b(\gamma_i) = \log(1 + \exp(\gamma_i)), c(t_i, \tau) = 0$, $\tau = 1$.

4.2.2 The generalized linear mixed model

As a result of the generalization to a non-linear response function, the generalized linear model can be applied to binary classification tasks by estimating the unknown $\boldsymbol{\beta}$ [1] [2], followed by non-linear transformation $h(\mathbf{f}_{N+1}^T\hat{\boldsymbol{\beta}})$, resulting in a probabilistic prediction for the assignment $T(\mathbf{x}_{N+1}) = c_1$. However, in presence of correlated data, prediction under the GLM typically results in low prediction accuracy. Higher accuracy can be obtained from an extension of the GLM, resulting from appending a vector of random effects to the linear predictor. The resulting model, referred to as a generalized linear mixed model (GLMM) (26), is capable of making use of information contained in a description of correlation, resulting in improved prediction accuracy if correlation is present.

In the GLMM, $\mathbf{t}$ is assumed to consist of conditionally independent observations t_i, given the (latent) y_i. Given y_i, t_i follow a distribution from an exponential family:

$p(\mathbf{t}|\mathbf{y}) = \prod_{i=1}^{N} p(t_i|y_i) = \prod_{i=1}^{N} \exp(\frac{t_i\gamma_i - b(\gamma_i)}{\tau} - c(t_i, \tau))$

The linear predictor $\boldsymbol{\eta} = \mathbf{F}\boldsymbol{\beta}$ is extended by a vector of random effects, interpreted as an incomplete realization of a Gaussian process $Y(\cdot)$:

$\boldsymbol{\eta} = \mathbf{F}\boldsymbol{\beta} + \mathbf{y}$

Finally, the structural assumption in the GLMM is

$E(\mathbf{t}|\mathbf{y}) = \boldsymbol{\mu} = h(\boldsymbol{\eta}) = h(\mathbf{F}\boldsymbol{\beta} + \mathbf{y})$, or, equivalently, $\mathbf{F}\boldsymbol{\beta} + \mathbf{y} = \boldsymbol{\eta} = g(\boldsymbol{\mu})$.

4.2.2.1 Prediction under the GLMM

Given a new data point $\mathbf{x}_{N+1}$ with known vector $\mathbf{f}_{N+1}$, prediction under the GLMM can be made by estimating the unknown parameters $\boldsymbol{\beta}$ and $\mathbf{y}$, followed by a non-linear transformation $h(\mathbf{f}_{N+1}^T\boldsymbol{\beta} + \mathbf{y})$ (substituting estimates for parameters), resulting in $p(T(\mathbf{x}_{N+1}) = c_1)$ for the assignment $T(\mathbf{x}_{N+1}) = c_1$.

As in the GLM, the unknown parameters are obtained by maximizing the marginal log likelihood of the model, given by

[1] In general, $\boldsymbol{\beta}$ is estimated by maximizing the marginal log likelihood of the model, resulting in a non-linear optimization problem, solved by IWLS, or Fisher scoring.

[2] see Appendix D for details

$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int \prod_{i=1}^{N} \exp(\frac{t_i\gamma_i - b(\gamma_i)}{\tau} - c(t_i, \tau))p(\mathbf{y})d\mathbf{y}$

with $p(\mathbf{y}) \sim N(\mathbf{y}|\mathbf{0}, \mathbf{C}_y)$,

with respect to $\boldsymbol{\beta}$ and $\mathbf{y}$.

In the general case (if the likelihood $p(\mathbf{t}|\mathbf{y})$ is not Gaussian), the integral cannot be evaluated analytically. Hence, application of the GLMM to classification tasks requires additional effort, with common approaches described below.

Numerical integration Numerical integration techniques approximate integrals which cannot be solved analytically. In general, these methods are based on re-expressing a function $f(\mathbf{x}) : \mathrm{I\!R}^N \rightarrow \mathrm{I\!R}_+$ as the product of a positive weight function $w(\mathbf{x}) : \mathrm{I\!R}^N \rightarrow \mathrm{I\!R}_+$ and a (real) function $g(\mathbf{x}) : \mathrm{I\!R}^N \rightarrow \mathrm{I\!R}$,
so that $f(\mathbf{x}) = w(\mathbf{x})g(\mathbf{x})$.

Gauss-Hermite quadrature In case when $\mathbf{x} = \boldsymbol{\theta}$, $p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, Gauss-Hermite quadrature can be applied to evaluate integrals of the form

$S(a(\boldsymbol{\theta})) = \int a(\boldsymbol{\theta})p(\mathbf{t}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$

Substituting $\boldsymbol{\theta} = \sqrt{2}\mathbf{L}\mathbf{z} + \boldsymbol{\mu}$ with the left Cholesky square root (the lower triangular factor) $\mathbf{L}$ [1] yields the expression

$S(a(\boldsymbol{\theta})) = \int a(\mathbf{z})p(\mathbf{t}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int a(\mathbf{z})p(\mathbf{t}|\mathbf{z})(2\boldsymbol{\pi})^{-\frac{N}{2}} \exp(-\mathbf{z}^T\mathbf{z})d\mathbf{z}$

with weight function $\exp(-\mathbf{z}^T\mathbf{z})$, which can be approximated using a Cartesian product rule:

$S(a(\boldsymbol{\theta})) = \int a(\mathbf{z})p(\mathbf{t}|\mathbf{z})(2\boldsymbol{\pi})^{-\frac{N}{2}} \exp(-z_1^2)\dots\exp(-z_N^2)dz_1\dots dz_N$
$\approx \sum_{i_1=1}^{k_1}(2\boldsymbol{\pi})^{-\frac{N}{2}}w_{i_1}^{(1)}\dots\sum_{i_N=1}^{k_N}(2\boldsymbol{\pi})^{-\frac{N}{2}}w_{i_N}^{(N)}a(z_{i_1}^{(1)},\dots,z_{i_N}^{(N)})p(\mathbf{t}|z_{i_1}^{(1)},\dots,z_{i_N}^{(N)})$

with $w_{i_r}^{(r)}$ denoting the weight of the Hermite polynomial $H_{k_r}(x)$ of degree k_r, evaluated at the i_r-th zero $x_{i_r}^{(r)}$, and $\mathbf{z} = (z_{i_1}^{(1)},\dots,z_{i_N}^{(N)})$, with $z_{i_r}^{(r)}$ denoting the

[1] obtained from the Cholesky decomposition $\mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}$, for a positive-definite matrix $\boldsymbol{\Sigma}$.

i_r-th zero of $H_{k_r}(z)$. [1]

Using the Cartesian product rule, an estimate for $\boldsymbol{\beta}$ can be obtained by maximization of the approximate marginal log likelihood $\log p(\mathbf{t})$,

$$\log p(\mathbf{t}) = \log S(1) = \int p(\mathbf{t}|\mathbf{z})(2\boldsymbol{\pi})^{-\frac{N}{2}} \exp\left(-\mathbf{z}^T\mathbf{z}\right)d\mathbf{z}$$
$$\approx \sum_{i_1=1}^{k_1} v_{i_1}^{(1)} \dots \sum_{i_N=1}^{k_N} v_{i_N}^{(N)} p(\mathbf{t}|z_{i_1}^{(1)}, \dots, z_{i_N}^{(N)}),$$

with $\mathbf{z} = \frac{1}{\sqrt{2}}\mathbf{L}^{-1}\mathbf{y}$,

given an approximation to the partial derivative $\frac{\partial \log(p(\mathbf{t}))}{\partial \boldsymbol{\beta}}$:

$$\frac{\partial \log(p(\mathbf{t}))}{\partial \boldsymbol{\beta}} = \frac{\int (\frac{\partial p(\mathbf{t}|\mathbf{y})}{\partial \boldsymbol{\beta}})p(\mathbf{y})d\mathbf{y}}{p(\mathbf{t})} = \frac{\int (\frac{1}{p(\mathbf{t}|\mathbf{y})}\frac{\partial p(\mathbf{t}|\mathbf{y})}{\partial \boldsymbol{\beta}})p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y}}{p(\mathbf{t})}$$
$$= \frac{\int \frac{\partial \log(p(\mathbf{t}|\mathbf{y}))}{\partial \boldsymbol{\beta}} p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y}}{p(\mathbf{t})}$$

Subsequently, $\hat{\boldsymbol{\beta}}$ can be substituted in the expression

$$E(\mathbf{y}|\mathbf{t}) = \frac{S(\mathbf{y})}{S(1)} \approx \frac{\sum_{i_1=1}^{k_1} v_{i_1}^{(1)} \dots \sum_{i_N=1}^{k_N} v_{i_N}^{(N)} \mathbf{z} p(\mathbf{t}|z_{i_1}^{(1)}, \dots, z_{i_N}^{(N)})}{p(\mathbf{t})},$$

yielding an estimate $E(\mathbf{y}|\mathbf{t}, \hat{\boldsymbol{\beta}})$.

Gauss-Hermite quadrature can be applied to classification tasks in the way described above. Unfortunately, the technique is limited with respect to the dimension of the integral, due to its computational complexity (in the order of k^N[2]) Hence, for large N (with N denoting the dimensionality of $\mathbf{y}$), other techniques must be applied.

Analytical approximations The computational complexity of numerical integration in case of large data sets suggests that numerical integration should be avoided. A different approach to estimation in the GLMM is based on an analytical approximation to the integrand by means of Laplace's approximation (44), resulting in a Gaussian approximation to the posterior $p(\boldsymbol{\beta}, \mathbf{y}|\mathbf{t})$, obtained by setting the posterior mean $E(\boldsymbol{\beta}, \mathbf{y}|\mathbf{t})$ to the posterior mode (obtained by maximization of $p(\mathbf{t}|\boldsymbol{\beta}, \mathbf{y})p(\mathbf{y})$) and the covariance matrix to the inverse of the Hessian

[1] The weights $w_{i_r}^{(r)}$ and nodes (quadrature points) $z_{i_r}^{(r)}$ are tabulated, e.g. in (1), p. 924.

[2] with $k = \max_i k_i$

$\mathbf{H}(\boldsymbol{\beta},\mathbf{y})$ (in case of Fisher Scoring, the Fisher matrix $\mathcal{I}(\boldsymbol{\beta},\mathbf{y})$) of $p(\mathbf{t}|\boldsymbol{\beta},\mathbf{y})p(\boldsymbol{\beta},\mathbf{y})$, evaluated at the mode.

Laplace's approximation The basic form of Laplace's approximation for evaluating the integral in the likelihood is based on a second-order Taylor series expansion around the maximum of $p(\mathbf{t}|\mathbf{y})p(\mathbf{y})$:

$$\log\int\exp(\log p(\mathbf{t}|\mathbf{y})+\log p(\mathbf{y}))=\log\int\exp(h(\mathbf{y}))d\mathbf{y}$$
$$\approx h(\mathbf{y}_0)+(\mathbf{y}-\mathbf{y}_0)\tfrac{\partial h(\mathbf{y})}{\partial\mathbf{y}}|_{\mathbf{y}=\mathbf{y_0}}+\tfrac{1}{2}(\mathbf{y}-\mathbf{y}_0)^T|\tfrac{\partial^2 h(\mathbf{y})}{\partial\mathbf{y}\partial\mathbf{y}^T}|_{\mathbf{y}=\mathbf{y_0}}(\mathbf{y}-\mathbf{y}_0),$$

where $\mathbf{y}_0$ is the solution to $\frac{\partial h(\mathbf{y})}{\partial\mathbf{y}}|_{\mathbf{y}=\mathbf{y_0}}=0$,

i.e., the mode of $\exp(h(\mathbf{y}))=\exp(\log p(\mathbf{t}|\mathbf{y})+\log p(\mathbf{y}))=p(\mathbf{t}|\mathbf{y})p(\mathbf{y})$.

Assuming a flat prior $p(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ (e.g., $p(\boldsymbol{\beta})\sim N(\boldsymbol{\beta}|\boldsymbol{\beta}_0,\mathbf{C}_\beta)$, with $\mathbf{C}_\beta^{-1}=\mathbf{0}$), and defining $\boldsymbol{\delta}^T=(\boldsymbol{\beta}^T,\mathbf{y}^T)$, so that the components of

$\frac{\partial\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial\boldsymbol{\delta}}=(\frac{\partial\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial\boldsymbol{\beta}},\frac{\partial\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial\mathbf{y}})$ are given by

$\frac{\partial\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial\boldsymbol{\beta}}\overset{\mathbf{C}_\beta^{-1}=\mathbf{0}}{=}\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{t}-\boldsymbol{\mu})$, and, by analogy,

$\frac{\partial\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial\mathbf{y}}=\frac{1}{\tau^2}\mathbf{W}\boldsymbol{\Delta}(\mathbf{t}-\boldsymbol{\mu})-\mathbf{C}_y^{-1}\mathbf{y}$,

the expression $\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}$ can be maximized with respect to $\boldsymbol{\delta}=(\boldsymbol{\beta},\mathbf{y})$. As in the case of the GLM, the expression is obtained through IWLS, e.g. Newton-Raphson, or Fisher Scoring (see Appendix D for details).

In order to evaluate the IWLS update step, it is necessary to obtain the Hessian $\mathbf{H}(\boldsymbol{\delta})$ or the Fisher information matrix $\mathcal{I}(\boldsymbol{\delta})$:

$$\frac{\partial^2\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T}=\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}(-1)\frac{\partial\mu}{\partial\boldsymbol{\beta}^T}+\frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\boldsymbol{\beta}^T}(\mathbf{t}-\boldsymbol{\mu})$$
$$=-\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\mathbf{F}+\frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\boldsymbol{\beta}^T}(\mathbf{t}-\boldsymbol{\mu}),$$

$$\frac{\partial^2\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\mathbf{y}^T}=(\frac{\partial^2\log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta})}{\partial\mathbf{y}\boldsymbol{\beta}^T})^T$$
$=\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}(-1)\frac{\partial\mu}{\partial\mathbf{y}^T}+\frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\mathbf{y}^T}(\mathbf{t}-\boldsymbol{\mu})=-\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}+\frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\mathbf{y}^T}(\mathbf{t}-\boldsymbol{\mu})$, and

$$\frac{\partial^2 \log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial \mathbf{y}\mathbf{y}^T} = \frac{\partial^2 h(\mathbf{y})}{\partial \mathbf{y}\partial \mathbf{y}^T} = -\frac{1}{\tau^2}\mathbf{W} + \frac{1}{\tau^2}\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial \mathbf{y}^T}(\mathbf{t}-\boldsymbol{\mu}) - \mathbf{C}_y^{-1}$$

Collecting terms, the Hessian matrix is

$$\mathbf{H}(\delta) = \begin{pmatrix} \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\beta}\mathbf{y}} \\ \mathbf{H}_{\mathbf{y}\boldsymbol{\beta}} & \mathbf{H}_{\mathbf{y}\mathbf{y}} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\mathbf{F} + \mathbf{A} & -\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W} + \mathbf{B} \\ (-\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W} + \mathbf{B})^T & -\frac{1}{\tau^2}\mathbf{W} + \mathbf{D} - \mathbf{C}_b^{-1} \end{pmatrix},$$

where $\mathbf{A} = \frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\boldsymbol{\beta}^T}(\mathbf{t}-\boldsymbol{\mu})$, $\mathbf{B} = \frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\mathbf{y}^T}(\mathbf{t}-\boldsymbol{\mu})$, and $\mathbf{D} = \frac{1}{\tau^2}\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\mathbf{y}^T}(\mathbf{t}-\boldsymbol{\mu})$.

Hence, the Fisher information matrix $\mathcal{I}(\boldsymbol{\delta})$ is

$$\mathcal{I}(\boldsymbol{\delta}) = \begin{pmatrix} -E(\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}) & -E(\mathbf{H}_{\boldsymbol{\beta}\mathbf{y}}) \\ -E(\mathbf{H}_{\mathbf{y}\boldsymbol{\beta}}) & -E(\mathbf{H}_{\mathbf{y}\mathbf{y}}) \end{pmatrix} \overset{\mathrm{E}(\mathbf{t}|\boldsymbol{\beta})=\boldsymbol{\mu}}{=} \begin{pmatrix} \frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\mathbf{F} & \frac{1}{\tau^2}\mathbf{F}^T\mathbf{W} \\ \frac{1}{\tau^2}\mathbf{W}\mathbf{F} & \frac{1}{\tau^2}\mathbf{W} + \mathbf{C}_y^{-1} \end{pmatrix}$$

and the posterior mode can be obtained using the update rule

$$\boldsymbol{\delta}^{(m+1)} = \boldsymbol{\delta}^{(m)} - \mathbf{H}(\boldsymbol{\delta}^{(m)})^{-1}\frac{\partial \log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial\boldsymbol{\delta}}, \text{ or}$$

$$\boldsymbol{\delta}^{(m+1)} = \boldsymbol{\delta}^{(m)} + \mathcal{I}(\boldsymbol{\delta}^{(m)})^{-1}\frac{\partial \log p(\mathbf{t}|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta}}{\partial\boldsymbol{\delta}}$$

, starting with an initial estimate $\boldsymbol{\delta}^{(0)}$.

In general, the justification for a Gaussian approximation to a posterior is that the true posterior will tend to a Gaussian as the number of data points increases (as a consequence of the central limit theorem). In case of Gaussian processes, the number of variables grows with the number of data points (hence, the argument does not apply directly). Hence, in context of spatial prediction, an alternative justification for a Gaussian approximation is suggested, based on the assumption of ergodicity. [1] Under the assumption of ergodicity, the Gaussian approximation improves with the size of the area [2], i.e., with $N \to \infty$.

One problem with Laplace's approximation under the GLMM is the presence of the $N \times P$ matrix $\mathbf{F}$, resulting in a $(P+N) \times (P+N)$ matrix to be inverted in each IWLS update step. For a polynomial trend of order p in D dimen-

[1] Loosely speaking, the ergodic hypothesis states that statistical averaging over realizations can be replaced by averaging over space .

[2] assuming fixed size for each $\mathbf{x}_i$

sions, the operation may become prohibitive, with time complexity in the order of $O((N+P)^3)$. Hence, in case of large/ complex data sets (with large N and $D > 3$, respectively), $\mathbf{F}\boldsymbol{\beta}$ may be dropped in favor of a model consisting only of $\mathbf{y}$, giving preference to a fully non-parametric simplification, as introduced below.

4.2.3 The GPM for classification

Application of the GLMM to the task of predicting the assignment $T(\mathbf{x}_{N+1}) = c_1$ (in case of binary classification) provides an alternative to indicator kriging, conforming to the assumption of a prior Gaussian process and a non-Gaussian noise process. Due to non-Gaussianity of the noise model, approximate inference must be performed to obtain estimates for unknown parameters, involving a variant of one of the approximations described in the preceding section.

In practice, both numerical integration and Laplace's approximation have been adopted. However, both techniques have their limitations, with respect to prediction accuracy (in case of Laplace's approximation) or scalability (in case of Gauss-Hermite quadrature). These techniques can be improved on by a particular variational approximation technique referred to as Expectation Propagation in machine learning, or by stochastic simulation techniques. Due to a close connection to Bayesian statistics, these methods are introduced in context of a Bayesian treatment of Gaussian process classification, as described below.

Gaussian process classification In the Bayesian approach to Gaussian process classification, a prior probability distribution is placed over an incomplete realization $\mathbf{y}$. As in the case of regression, it is assumed that $Y(\cdot)$ is a latent Gaussian process with zero mean and covariance function reflecting assumptions about characteristics of functions involved in inference:

$p(\mathbf{y}) = N(\mathbf{y}|\mathbf{0}, \mathbf{C})$

For inference, the prior is combined with a likelihood function, reflecting assumptions about the type of the (non-Gaussian) noise process $T_Y(\cdot)$. In case of binary classification, the likelihood is given by a product of Bernoulli distributions, with π_i given by the logistic function $\pi_i = \pi_i(y_i) = \frac{1}{\exp(-y_i)}$:

$p(\mathbf{t} = \mathbf{1}|\mathbf{y}) = \prod_{i=1}^{N} p(t_i = 1|\pi_i)$

$= \prod_{i=1}^{N} \binom{1}{1} \exp(1 \log(\pi_i) + (1-1)\log(1-\pi_i) + \log(\binom{1}{1}))) = \prod_{i=1}^{N} \pi_i,$

assuming that t_i is conditionally independent of other observations t_j given the latent y_i, with $i \neq j$. As an alternative, the likelihood can chosen to be the cumulative distribution function of a standard normal distribution $\Phi(t_i y_i)$:

$p(\mathbf{t} = \mathbf{1}|\mathbf{y}) = \prod_{i=1}^{N} \Phi(t_i y_i),$

with $\pi_i = \Phi(t_i y_i) = \int_{-\infty}^{t_i y_i} N(x|0,1)dx,$

referred to as probit likelihood.

In the Bayesian approach to (binary) Gaussian process classification, inference is divided into two steps: First, the distribution of the latent $Y(\mathbf{x}_{N+1})$ at data point $\mathbf{x}_{N+1}$ given the observations $\mathbf{t}$, given by

$p(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1}) = \int p(Y(\mathbf{x}_{N+1})|\mathbf{x}_{N+1}, \mathbf{y})p(\mathbf{y}|\mathbf{t})d\mathbf{y},$

with $p(\mathbf{y}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{t})},$

is evaluated. Subsequently, the distribution over $Y(\mathbf{x}_{N+1})$ is used to obtain a probabilistic prediction

$p(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$
$= \int p(T(\mathbf{x}_{N+1}) = 1|Y(\mathbf{x}_{N+1}))p(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1})dY(\mathbf{x}_{N+1})$

In regression, inference can be performed analytically, making use of properties of the Gaussian. In classification, the non-Gaussian likelihood $p(\mathbf{t}|\mathbf{y})$ in the posterior $p(\mathbf{y}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{t})}$ makes the (in general, high-dimensional) integral $\int p(Y(\mathbf{x}_{N+1})|\mathbf{x}_{N+1}, \mathbf{y})p(\mathbf{y}|\mathbf{t})d\mathbf{y}$ analytically intractable. Consequently, to perform inference, approximations must be applied, with suitable techniques introduced in the following sections.

4.2.3.1 Analytical approximations

Laplace's approximation Laplace's method for the GPM, introduced in (47), involves a Gaussian approximation $q(\mathbf{y}|\mathbf{t})$ to the posterior $p(\mathbf{y}|\mathbf{t})$. As in Laplace's method for the GLMM, the approximation is obtained through a second order Taylor expansion of $\log p(\mathbf{y}|\mathbf{t})$ around the maximum of the posterior:

$$q(\mathbf{y}|\mathbf{t}) = N(\mathbf{y}|\hat{\mathbf{y}}, -\mathbf{H}^{-1}) \propto \exp(-\tfrac{1}{2}(\mathbf{y}-\hat{\mathbf{y}})^T(-\mathbf{H})(\mathbf{y}-\hat{\mathbf{y}})),$$

where $\hat{\mathbf{y}} = \arg\max_y p(\mathbf{y}|\mathbf{t})$, and $\mathbf{H} = \frac{\partial^2 \log p(\mathbf{y}|\mathbf{t})}{\partial \mathbf{y} \partial \mathbf{y}^T}|_{\mathbf{y}=\hat{\mathbf{y}}}$ denotes the Hessian matrix of $-\log p(\mathbf{y}|\mathbf{t})$, evaluated at $\hat{\mathbf{y}}$.

Substituting the expression for the Gaussian for $p(\mathbf{y})$, the log of the un-normalized posterior $\Psi(\mathbf{y}) = p(\mathbf{t}|\mathbf{y})p(\mathbf{y})$ is

$$\Psi(\mathbf{y}) = \log p(\mathbf{t}|\mathbf{y}) + \log p(\mathbf{y}) = \log p(\mathbf{t}|\mathbf{y}) - \tfrac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} - \log|\mathbf{C}| - \tfrac{N}{2}\log(2\pi)$$

In order to find $\hat{\mathbf{y}}$, the IWLS algorithm is applied (see Appendix D for details), with the Newton-Raphson update step

$$\mathbf{y}^{(m+1)} = \mathbf{y}^{(m)} - \mathbf{H}_{\Psi}(\mathbf{y})^{-1}\frac{\partial \log p(\mathbf{t}|\mathbf{y})p(\mathbf{y})}{\partial \mathbf{y}}|_{\mathbf{y}=\mathbf{y}^{(m)}}$$
$$= \mathbf{y}^{(m)} - (\frac{\partial^2 \log p(\mathbf{t}|\mathbf{y})p(\mathbf{y})}{\partial \mathbf{y}\partial \mathbf{y}^T}|_{\mathbf{y}=\mathbf{y}^{(m)}})^{-1}\frac{\partial \log p(\mathbf{t}|\mathbf{y})p(\mathbf{y})}{\partial \mathbf{y}}|_{\mathbf{y}=\mathbf{y}^{(m)}}$$
$$= (\mathbf{C}^{-1}+\mathbf{W})^{-1}(\frac{\partial \log p(\mathbf{t}|\mathbf{y})}{\partial \mathbf{y}}|_{\mathbf{y}=\mathbf{y}^{(m)}} - \mathbf{C}^{-1}\mathbf{y}),$$

where

$$\frac{\partial \log p(\mathbf{t}|\mathbf{y})p(\mathbf{y})}{\partial \mathbf{y}} = \frac{\partial \log p(\mathbf{t}|\mathbf{y})}{\partial \mathbf{y}} - \mathbf{C}^{-1}\mathbf{y},\ \frac{\partial^2 \log p(\mathbf{t}|\mathbf{y})p(\mathbf{y})}{\partial \mathbf{y}\partial \mathbf{y}^T} = -\mathbf{W} - \mathbf{C}^{-1},$$

$$\mathbf{W} = \begin{pmatrix} \frac{\partial^2 \log p(t_1|y_1)}{\partial y_1^2} & 0 & \dots & 0 \\ 0 & \frac{\partial^2 \log p(t_2|y_2)}{\partial y_2^2} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \frac{\partial^2 \log p(t_N|y_N)}{\partial y_N^2} \end{pmatrix}$$

At the maximum of $\Psi(\mathbf{y})$, Laplace's approximation to $p(\mathbf{y}|\mathbf{t})$ results in a Gaussian with mean $\hat{\mathbf{y}}$ and covariance matrix $-\mathbf{H}_{\Psi}^{-1} = (\mathbf{C}^{-1}+\mathbf{W})^{-1}$:

$$p(\mathbf{y}|\mathbf{t}) \approx q(\mathbf{y}|\mathbf{t}) = N(\mathbf{y}|\hat{\mathbf{y}}, -\mathbf{H}_{\Psi}^{-1}) = N(\mathbf{y}|\hat{\mathbf{y}}, (-\frac{\partial^2 \log p(\mathbf{y}|\mathbf{t})}{\partial \mathbf{y}\partial \mathbf{y}^T}|_{\mathbf{y}=\hat{\mathbf{y}}})^{-1}) = N(\mathbf{y}|\boldsymbol{\mu} = \hat{\mathbf{y}}, \boldsymbol{\Sigma} = (\mathbf{C}^{-1}+\mathbf{W})^{-1})$$

Given $q(\mathbf{y}|\mathbf{t})$, the first inference step is performed making use of the result for the conditional Gaussian distribution $p(\mathbf{x}_1|\mathbf{x}_2)$, given the joint Gaussian distribution $p(\mathbf{x}_1, \mathbf{x}_2)$ (see Appendix A for details):

$$p(Y(\mathbf{x}_{N+1})|\mathbf{x}_{N+1}, \mathbf{y}) = N(Y(\mathbf{x}_{N+1})|\mathbf{c}\mathbf{C}^{-1}\mathbf{y}, c - \mathbf{c}\mathbf{C}^{-1}\mathbf{c}^T),$$

with $c = var(Y(\mathbf{x}_{N+1}))$, and $\mathbf{c} = cov(T(\mathbf{x}_{N+1}), \mathbf{t})$.

Making use of the result for the marginal distribution $p(\mathbf{y})$ given a marginal distribution $p(\mathbf{x})$ and a conditional distribution $p(\mathbf{y}|\mathbf{x})$, (see Appendix A for details), the expression is combined with $q(\mathbf{y}|\mathbf{t})$:

$$q(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1}) = N(Y(\mathbf{x}_{N+1})|\mathbf{c}\mathbf{C}^{-1}\hat{\mathbf{y}}, c - \mathbf{c}(\mathbf{C} + \mathbf{W}^{-1})^{-1}\mathbf{c}^T),$$

Given above expression, Laplace's method for binary classification results in an approximation to $p(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$:

$$p(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1}) \approx q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$$
$$= \int p(T(\mathbf{x}_{N+1}) = 1|Y(\mathbf{x}_{N+1}))q(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1})dY(\mathbf{x}_{N+1})$$

Given this expression $q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$, a probability for the assignment $T(\mathbf{x}_{N+1}) = 1$ given $\mathbf{t}$ can be obtained.

If p$(T(\mathbf{x}_{N+1}) = 1|Y(\mathbf{x}_{N+1})) = \Phi(t_i y_i)$, $q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$ can be evaluated analytically:

$$q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$$
$$= \int \Phi(\tfrac{x-m}{v})N(x|\mu, \sigma^2)dx = \tfrac{1}{(2\pi(v^2+\sigma^2))^{\frac{1}{2}}} \int_{-\infty}^{\mu-m} \exp(-\tfrac{z^2}{2(v^2+\sigma^2)})dz = \Phi(\tfrac{\mu-m}{(v^2+\sigma^2)^{\frac{1}{2}}})$$
$$\overset{\mu=\mathbf{c}\mathbf{C}^{-1}\hat{\mathbf{y}},\sigma^2=c-\mathbf{c}(\mathbf{C}+\mathbf{W}^{-1})^{-1}\mathbf{c}^T}{=} \Phi(\tfrac{\mathbf{c}\mathbf{C}^{-1}\hat{\mathbf{y}}}{\sqrt{(1+c-\mathbf{c}(\mathbf{C}+\mathbf{W}^{-1})^{-1}\mathbf{c}^T)}})$$

In case $p(T(\mathbf{x}_{N+1}) = 1|Y(\mathbf{x}_{N+1})) = \frac{1}{\exp(-y_i)}$, the integral in $q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$ is analytically intractable , and the expression must be approximated. In contrast to the integral in $p(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1})$, the integral in $q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$ is one-dimensional, and numerical integration is feasible.

In context of Gaussian process classification, application of Laplace's method has several advantages, including applicability to large/ complex classification problems and reliable convergence to the (global) maximum of the (un-normalized) posterior $p(\mathbf{t}|\mathbf{y})p(\mathbf{y})$ (assuming the likelihood $p(\mathbf{t}|\mathbf{y})$ is log concave [1]). Unfortunately, a problem with the Laplace approximation is that the Hessian matrix (evaluated at the mode $\hat{\mathbf{y}}$) may give a poor approximation to the true shape of the posterior (e.g. if the true posterior is skewed). In such a case, a better approximation to the $q(\mathbf{y}|\mathbf{t})$ may be obtained by means of an alternative approach, involving a (local) Gaussian approximation to the contribution of each data point to the likelihood, as described below.

Expectation Propagation The Expectation Propagation (EP) technique, introduced in (28), is a general analytical approximation scheme applicable to a range of taks, including regression, probabilistic classification, and count type regression. In this section, the application to (binary) classification is described, with likelihood function $p(\mathbf{t} = \mathbf{1}|\mathbf{y}) = \prod_{i=1}^{N} p(t_i = 1|\pi_i)$ given by $\prod_{i=1}^{N} p(t_i = 1|\pi_i) = \prod_{i=1}^{N} \Phi(t_i y_i)$.

In a nutshell, Expectation Propagation for binary classification involves a local Gaussian approximation to the contribution of each data point to the likelihood, in the form of an (un-normalized) Gaussian in the latent y_i:

$$p(t_i = 1|\pi_i) = p(t_i = 1|\pi_i(y_i)) \approx g_i(y_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \tilde{Z}_i N(y_i|\tilde{\mu}_i, \tilde{\sigma}_i^2),$$

with site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$, and $\tilde{\sigma}_i^2$, corresponding to the 0-th, first, and second moment of the (normalized) Gaussian $N(y_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)$.

The product of the (independent) local likelihoods g_i is

$$\prod_{i=1}^{N} g_i(y_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \prod_{i=1}^{N} \tilde{Z}_i,$$

with $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_N)^T$,

and $\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \tilde{\sigma}_1^2 & 0 & \ldots & 0 \\ 0 & \tilde{\sigma}_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & \ldots & \tilde{\sigma}_N^2 \end{pmatrix}$, so that

[1] This is the case for the the logistic and the probit response function.

$p(\mathbf{t} = \mathbf{1}|\mathbf{y}) = \prod_{i=1}^{N} p(t_i = 1|\pi_i) = N(\mathbf{y}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \prod_{i=1}^{N} \tilde{Z}_i.$

In Expectation Propagation, the posterior distribution $p(\mathbf{y}|\mathbf{t})$ is approximated by a Gaussian of the form

$q(\mathbf{y}|\mathbf{t}) = \frac{1}{Z_{EP}} \prod_{i=1}^{N} g_i(y_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) p(\mathbf{y}) = N(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$

with $\boldsymbol{\mu} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}$, and $\boldsymbol{\Sigma} = (\mathbf{C}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1}$, resulting from

$N(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{Z_{EP}} N(\mathbf{y}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \prod_{i=1}^{N} \tilde{Z}_i N(\mathbf{y}|\mathbf{0}, \mathbf{C})$
$= N(\mathbf{y}|\boldsymbol{\mu} = \boldsymbol{\Sigma}(\mathbf{C}^{-1}\mathbf{0} + \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}), \boldsymbol{\Sigma} = (\mathbf{C}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1})$ (see Appendix A for details),

and $Z_{EP} = q(\mathbf{t})$, $p(\mathbf{t}) \approx q(\mathbf{t})$ denoting the EP approximation to the marginal likelihood $p(\mathbf{t})$.

In order to obtain $q(\mathbf{y}|\mathbf{t})$, the site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ of the local approximations g_i must be found. In EP, this is done as part of an iterative procedure, in which $\tilde{Z}_i$, $\tilde{\mu}_i$, and $\tilde{\sigma}_i^2$ are updated sequentially. This is done by iterating the following steps (for $1 \leq i \leq N$), until convergence:

1. Starting with a current approximate posterior $q(\mathbf{y}|\mathbf{t})$, the current g_i is left out, resulting in an approximate Gaussian cavity distribution $q_{-i}(y_i) = N(y_i|\mu_{-i}, \sigma_{-i}^2) \propto \int \prod_{j \neq i} g_j(y_j|\tilde{Z}_j, \tilde{\mu}_j, \tilde{\sigma}_j^2) p(\mathbf{y}) dy_j$, so that

   $g_i(y_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) q_{-i}(y_i) = \tilde{Z}_i N(y_i|\tilde{\mu}_i, \tilde{\sigma}_i^2) N(y_i|\mu_{-i}, \sigma_{-i}^2) = N(y_i|\mu_i, \sigma_i^2 = \boldsymbol{\Sigma}_{ii}).$

2. In the second step, the cavity distribution is combined with the exact likelihood $p(t_i = 1|y_i)$, resulting in a non-Gaussian marginal $\int q_{-i}(y_i) p(t_i = 1|y_i) dy_i$.

3. In the third step, an (un-normalized) Gaussian posterior marginal $\hat{q}(y_i)$ approximating $\int q_{-i}(y_i) p(t_i = 1|y_i) dy_i$ is found by minimizing the Kullback-Leibler divergence $KL(p(y_i)\|q(y_i)) = KL(\int q_{-i}(y_i) p(t_i = 1|y_i) dy_i \| q(y_i)) = -\int q_{-i}(y_i) p(t_i = 1|y_i) dy_i \log \frac{q(y_i)}{\int q_{-i}(y_i) p(t_i=1|y_i) dy_i}$:

$\hat{q}(y_i) = \hat{Z}_i N(y_i|\hat{\mu}_i, \hat{\sigma}_i^2)$
$= \arg\min_{q(y_i)} KL(\int q_{-i}(y_i)p(t_i = 1|y_i)dy_i \| q(y_i))$

4. In the fourth step, a local approximation g_i is found, so that

$$g_i(y_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)q_{-i}(y_i) = \tilde{Z}_i N(y_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)N(y_i|\mu_{-i}, \sigma_{-i}^2) = \hat{Z}_i N(y_i|\hat{\mu}_i, \hat{\sigma}_i^2).$$

5. Finally, $q(\mathbf{y}|\mathbf{t})$ is updated to include the g_i (the site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$, and $\tilde{\sigma}_i^2$) obtained in the last step.

In more detail, g_i are optimized sequentially, using the approximations obtained so far for all g_j, $j \neq i$. Typically, this requires several passes over the data, since the update of a g_i potentially influences all approximate marginal posteriors.

Due to minimization of the Kullback-Leibler divergence $KL(p(y_i)\|q(y_i))$, Expectation Propagation has been referred to as a variational approximation technique in machine learning. For a Gaussian $q(y_i)$, the $\hat{q}(y_i)$ minimizing $KL(p(y_i)\|q(y_i))$ is a Gaussian whose first and second moments match the first and second moments of $p(y_i)$. [1]

Based on the EP approximation $q(\mathbf{y}|\mathbf{t})$, prediction for a new data point $\mathbf{x}_{N+1}$ can be made as in the case of Laplace's approximation, with the EP approximation to $p(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1})$ given by:

$$q(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1}) = N(Y(\mathbf{x}_{N+1})|\mathbf{c}(\mathbf{C} + \tilde{\mathbf{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}}, c - \mathbf{c}(\mathbf{C} + \tilde{\mathbf{\Sigma}})^{-1}\mathbf{c}^T),$$

with $c = var(Y(\mathbf{x}_{N+1}))$ and $\mathbf{c} = cov(T(\mathbf{x}_{N+1}), \mathbf{t})$.

Given above expression, EP for binary classification results in an approximation to $p(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$, given by

$q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$
$= \int p(T(\mathbf{x}_{N+1}) = 1|Y(\mathbf{x}_{N+1}))q(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1})dY(\mathbf{x}_{N+1})$

From above expression, a probability for the assignment $T(\mathbf{x}_{N+1}) = 1$ can be obtained, resulting in the prediction

[1] As $\hat{q}(y_i)$ is un-normalized, the 0-th moment (the normalizing constant) $\hat{Z}_i$ of $\hat{q}(y_i)$ is equated to $\int N(y_i|\mu_{-i}, \sigma_{-i}^2)\Phi(\frac{t_i y_i - 0}{1})dy_i = \Phi(z_i)$, with $z_i = \frac{t_i \mu_{-i}}{\sqrt{1+\sigma_{-i}^2}}$.

$E_p(p(T(\mathbf{x}_{N+1}) = 1|Y(\mathbf{x}_{N+1}))) \approx E_q(p(T(\mathbf{x}_{N+1}) = 1|Y(\mathbf{x}_{N+1}))) = q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1})$
$= \int \Phi(\frac{x-m}{v}) N(x|\mu, \sigma^2) dx$
$\overset{\mu=\mathbf{c}(\mathbf{C}+\tilde{\mathbf{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}}, \sigma^2=c-\mathbf{c}(\mathbf{C}+\tilde{\mathbf{\Sigma}})^{-1}\mathbf{c}^T}{=} \Phi(\frac{\mathbf{c}(\mathbf{C}+\tilde{\mathbf{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}},}{\sqrt{(1+c-\mathbf{c}(\mathbf{C}+\tilde{\mathbf{\Sigma}})^{-1}\mathbf{c}^T)}})$

4.2.3.2 Markov Chain Monte Carlo

In many probabilistic models, Bayesian inference, involving the evaluation of high-dimensional integrals, is intractable, and approximate inference techniques must be applied. In case of Gaussian process models for classification, deterministic techniques involving a Gaussian approximation to the posterior $p(\mathbf{y}|\mathbf{t})$ can be used, and inference can be performed analytically, as described in the previous sections. An alternative to this approach is provided by stochastic approximation techniques. In general, these techniques are based on the idea that the expectation of a function $f(\boldsymbol{\theta})$ with respect to a (un-normalized) distribution $p(\boldsymbol{\theta})$[1] can be approximated by averaging over a finite set of samples $\boldsymbol{\theta}_l$ (with $l = 1, \ldots, L$) from $p(\boldsymbol{\theta})$:

$E(f(\boldsymbol{\theta})) = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \hat{f}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{l=1}^{L} f(\boldsymbol{\theta}_l)$

An advantage of stochastic approximation techniques is that, in principle, the accuracy of $\hat{f}(\boldsymbol{\theta})$ does not depend on the dimensionality of $\boldsymbol{\theta}$. In general, a problem with the approach is that obtaining a set of independent samples from $p(\boldsymbol{\theta})$ can be difficult (depending on the form of $p(\boldsymbol{\theta})$). This problem is addressed by a class of general and powerful stochastic approximation techniques, referred to as Markov Chain Monte Carlo (MCMC) techniques (13), (22), sampling in a way such that the distribution $p(\boldsymbol{\theta})$ is an invariant distribution of a Markov chain [2] in the space of $\boldsymbol{\theta}$:

$p(\boldsymbol{\theta}) = \sum_{\boldsymbol{\theta}'} z(\boldsymbol{\theta}', \boldsymbol{\theta}) p(\boldsymbol{\theta}'),$

[1] In general, $p(\boldsymbol{\theta})$ must be known up to a normalization constant.

[2] A Markov chain of order n (with $n \in \mathbb{N}_0$) is a collection of random variables $\mathbf{X}_m$ ($m = 1, \ldots, M$, with $M \in \mathbb{N}$), with the property that

$p(\mathbf{X}_m = \boldsymbol{\theta}_m|\mathbf{X}_0 = \boldsymbol{\theta}_0, \mathbf{X}_1 = \boldsymbol{\theta}_1, \ldots, \mathbf{X}_{m-1} = \boldsymbol{\theta}_{m-1}) = p(\mathbf{X}_m = \boldsymbol{\theta}_m|\mathbf{X}_{m-n} = \boldsymbol{\theta}_{m-n}, \ldots, \mathbf{X}_{m-1} = \boldsymbol{\theta}_{m-1}),$

for $n \leq m$.

with $z(\boldsymbol{\theta}', \boldsymbol{\theta}) = z(\boldsymbol{\theta}^m, \boldsymbol{\theta}^{m+1})$ denoting the transition probability from state $\boldsymbol{\theta}^m$ to $\boldsymbol{\theta}^{m+1}$.

In this approach, the proposal distribution $p_m(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^m)$ used to draw the next sample depends on the current state $\boldsymbol{\theta}^m$, and the sequence of samples $\boldsymbol{\theta}^0$, $\boldsymbol{\theta}^1$, $\boldsymbol{\theta}^2$, ..., has the Markov property:

$p(\boldsymbol{\theta}^{m+1}|\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^m) = p(\boldsymbol{\theta}^{m+1}|\boldsymbol{\theta}^m)$

Assuming that the Markov chain is ergodic, i.e., that for $m \to \infty$, $p(\boldsymbol{\theta}^m)$ converges to $p(\boldsymbol{\theta})$ (irrespective of the initial state $\boldsymbol{\theta}^0$), $\boldsymbol{\theta}^m$ becomes an approximately independent sample of $p(\boldsymbol{\theta})$. In practice, the chain is generated for a finite length L and the sequence $\boldsymbol{\theta}^0$, $\boldsymbol{\theta}^1$, $\boldsymbol{\theta}^2$, ..., $\boldsymbol{\theta}^L$ is interpreted as a set of samples drawn from $p(\boldsymbol{\theta})$. The procedure is continued until enough samples are obtained, i.e., so that the expectation of $f(\boldsymbol{\theta})$ with respect to $p(\boldsymbol{\theta})$ is approximated well by $\hat{f}(\boldsymbol{\theta})$.

One problem in MCMC is the choice of the proposal distribution used to draw the next sample, given the current state $\boldsymbol{\theta}^m$. In general, the proposal distribution is specified together with a rule indicating if a candidate sample $\tilde{\boldsymbol{\theta}}$ drawn from $p_m(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^m)$ is accepted, so that $\boldsymbol{\theta}^{m+1} := \tilde{\boldsymbol{\theta}}$ (i.e., a change of state occurs) or not, which constitutes a MCMC procedure.

Metropolis-Hastings A basic Markov Chain Monte Carlo technique is the Metropolis-Hastings sampling method (27), (16). In the Metropolis-Hastings procedure, a candidate sample $\tilde{\boldsymbol{\theta}}$ is drawn from a proposal distribution (depending on the current state $\boldsymbol{\theta}^m$). The candidate sample is accepted with probability $p_z(\tilde{\boldsymbol{\theta}})$, where

$p_z(\tilde{\boldsymbol{\theta}}) = \min\,(1, \frac{p(\tilde{\boldsymbol{\theta}})p_m(\boldsymbol{\theta}^m|\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}^m)p_m(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^m)})$

Given $p_z(\tilde{\boldsymbol{\theta}})$, the candidate sample is accepted (so that $\boldsymbol{\theta}^{m+1} := \tilde{\boldsymbol{\theta}}$) if $u \sim U(u|0,1) \leq p_z(\tilde{\boldsymbol{\theta}})$,

with $U(u|0,1)$ denoting the uniform distribution on the interval $[0,1]$, and rejected otherwise.

Gibbs sampling Markov Chain Monte Carlo sampling procedures for a particular $p(\boldsymbol{\theta})$ can be implemented using different proposal distributions. In a more refined proposal strategy, $\boldsymbol{\theta}$ can be divided in R components, so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)^T$. Then, instead of updating the state $\boldsymbol{\theta}$, each component $\boldsymbol{\theta}_r$ (with $r = 1, \ldots, R) \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R\}$ is updated in turn. This strategy is applied in Gibbs sampling, which can be seen as a special case of the Metropolis-Hastings procedure.

For a $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$, each step of Gibbs sampling involves replacing the value of a $\boldsymbol{\theta}_r$ by a value drawn from the distribution of $\boldsymbol{\theta}_r$ conditioned on the values of the remaining variables:

$$\boldsymbol{\theta}_r^{m+1} = \tilde{\boldsymbol{\theta}}_r,$$

with $\tilde{\boldsymbol{\theta}}_r \sim p(\boldsymbol{\theta}_r | \boldsymbol{\theta}^m_{-r})$

Gibbs sampling can be seen as a special case of Metropolis-Hastings in which the proposed states are always accepted. This can be seen by substituting the proposal distribution $p(\boldsymbol{\theta}_r | \boldsymbol{\theta}^m_{-r})$ in the Metropolis-Hastings rule:

$$p_z(\tilde{\boldsymbol{\theta}}) = \frac{p(\tilde{\boldsymbol{\theta}}_r, \boldsymbol{\theta}^m_{-r}) p_m(\boldsymbol{\theta}^m | \tilde{\boldsymbol{\theta}}_r, \boldsymbol{\theta}^m_{-r})}{p(\boldsymbol{\theta}^m) \quad p_m(\tilde{\boldsymbol{\theta}}_r, \boldsymbol{\theta}^m_{-r} | \boldsymbol{\theta}^m)} = 1,$$

with $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_r, \boldsymbol{\theta}^m_{-r})$.

A practical problem with Gibbs sampling is that dependencies between $\boldsymbol{\theta}_r$ are not taken into account. In effect, the chain may take long to explore the support of $p(\boldsymbol{\theta})$.

Hybrid Monte Carlo One problem of the Metropolis-Hastings procedure is sensitivity to step size. If the step size in a particular direction j (corresponding to a component θ_j) is too small, the procedure will take long to explore the support of $p(\boldsymbol{\theta})$. Conversely, if the step size is too large, the acceptance rate will be low, resulting in an inefficient procedure.

The technique of Hybrid Monte Carlo (9) provides an adaptive step size which is adjusted to match the characteristics of $p(\boldsymbol{\theta})$. In short, the technique is based on evaluating partial derivatives of $p(\boldsymbol{\theta})$ with respect to θ_j, which provide information about directions in which regions of higher density can be found.

Conceptually, Hybrid Monte Carlo can be thought of as a simulation of a fictitious physical system evolving in continuous time τ. In this system, the state $\boldsymbol{\theta}$ is interpreted as the location of particles with momentum $\mathbf{q}$. New states are proposed using a procedure which can be understood as a discrete simulation of Hamiltonian dynamics.

In order to simulate the system, the potential and kinetic energies have to be defined. The potential energy $E(\boldsymbol{\theta})$ is set to $-\log p(\boldsymbol{\theta})$. Then, the energy of the system (referred to as the Hamiltonian $H(\boldsymbol{\theta}, \mathbf{q})$) is the sum of potential energy $E(\boldsymbol{\theta})$ and kinetic energy $K(\mathbf{q})$, with

$$H(\boldsymbol{\theta}, \mathbf{q}) = E(\boldsymbol{\theta}) + K(\mathbf{q}) = -\log p(\boldsymbol{\theta}) + \tfrac{1}{2}\|\mathbf{q}\|^2$$

The key idea in Hybrid Monte Carlo is that candidate samples in a Metropolis-Hastings sampler are obtained by discrete simulation of Hamiltonian dynamics. The simulation is discretised using the so-called leapfrog method, which minimizes the impact of errors introduced in a numerical integration of the Hamiltonian equations:

$$\tfrac{d\boldsymbol{\theta}}{d\tau} = \tfrac{\partial H(\boldsymbol{\theta},\mathbf{q})}{\partial \mathbf{q}} = \mathbf{q},$$

$$\tfrac{d\mathbf{q}}{d\tau} = -\tfrac{\partial H(\boldsymbol{\theta},\mathbf{q})}{\partial \boldsymbol{\theta}} = \tfrac{\partial \log p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and ensures that $p(\boldsymbol{\theta}, \mathbf{q}) \propto \exp\left(-H(\boldsymbol{\theta}, \mathbf{q})\right) = p(\boldsymbol{\theta}) \exp\left(-\tfrac{1}{2}\|\mathbf{q}\|^2\right)$ is invariant.

In summary, Hybrid Monte Carlo involves alternating between a series of leapfrog updates and resampling of $\mathbf{q}$. After each application of the leapfrog updates, the resulting candidate state is accepted or rejected according to the Metropolis-Hastings rule based on the value of $H(\boldsymbol{\theta}, \mathbf{q})$:

$p_z(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{q}}) = \min\left(1, \exp\left(H(\boldsymbol{\theta}^m, \mathbf{q}^m) - H(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{q}})\right)\right)$

Unlike the basic Metropolis-Hastings procedure, Hybrid Monte Carlo is able to make use of information from the gradient $\frac{\partial \log p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. In general, this makes the Hybrid Monte Carlo procedure more efficient in exploring the support of $p(\boldsymbol{\theta})$.

Hybrid Monte Carlo for the GPM Hybrid Monte Carlo can be used to perform approximate inference in the GPM. In case of Gaussian process classification, $\mathbf{y}_l$ can be drawn from $p(\mathbf{y}[,\mathbf{t}]) = p(\mathbf{t}|\mathbf{y})p(\mathbf{y})$ (assuming differentiability of likelihood $p(\mathbf{t}|\mathbf{y})$ and prior $p(\mathbf{y})$), so that

$E(\mathbf{y}) = \int \mathbf{y} p(\mathbf{y}[,\mathbf{t}]) d\mathbf{y} \approx \hat{\mathbf{y}} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{y}_l$, and

$cov(\mathbf{y}) = \int (\mathbf{y} - E(\mathbf{y}))(\mathbf{y} - E(\mathbf{y}))^T p(\mathbf{y}[,\mathbf{t}]) d\mathbf{y}$
$\approx \hat{cov}(\mathbf{y}) = \frac{1}{L}\sum_{l=1}^{L}(\mathbf{y}_l - \hat{\mathbf{y}})(\mathbf{y}_l - \hat{\mathbf{y}})^T$

Hyperparameter sampling When applying a MCMC sampling procedure, it is straightforward to perform inference over all unknown parameters. In case of the GPM, this includes the vector of parameters of the covariance (denoted $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)^T$). Hybrid Monte Carlo can be used to draw $\boldsymbol{\theta}_l = (\mathbf{y}, \boldsymbol{\psi})_l^T$ from the distribution $p(\mathbf{y}, \boldsymbol{\psi}|\mathbf{t}) \propto p(\mathbf{y}, \boldsymbol{\psi}[,\mathbf{t}]) = p(\mathbf{t}|\mathbf{y})p(\mathbf{y}|\boldsymbol{\psi})p(\boldsymbol{\psi})$, or, more generally,

$p(\mathbf{y}, \boldsymbol{\psi}|\mathbf{t}, \boldsymbol{\xi}) \propto p(\mathbf{y}, \boldsymbol{\psi}[,\mathbf{t}]|\boldsymbol{\xi}) = p(\mathbf{t}|\mathbf{y})p(\mathbf{y}|\boldsymbol{\psi})p(\boldsymbol{\psi}|\boldsymbol{\xi})$,

introducing a hyperprior $p(\boldsymbol{\psi}|\boldsymbol{\xi})$ for hyperparameter $\boldsymbol{\psi}$.

Substituting expressions, $\boldsymbol{\theta}_l$ can be sampled from the un-normalized log posterior, given by

$\log p(\mathbf{y}, \boldsymbol{\psi}[,\mathbf{t}]|\boldsymbol{\xi}) = \log p(\mathbf{t}|\mathbf{y}) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} + \log p(\boldsymbol{\psi}|\boldsymbol{\xi})$,

with $\nabla p(\mathbf{y}, \boldsymbol{\psi}[,\mathbf{t}]|\boldsymbol{\xi}) = \nabla \log p(\mathbf{t}|\mathbf{y}) - \mathbf{C}^{-1}\mathbf{y}$ (with $\nabla \log p(\mathbf{t}|\mathbf{y})$ depending on the choice of the likelihood function),

and $\frac{\partial \log p(\mathbf{y}, \boldsymbol{\psi}[,\mathbf{t}]|\boldsymbol{\xi})}{\partial \psi_j} = -\frac{1}{2}tr(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \psi_j}) + \frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \psi_j}\mathbf{C}^{-1}\mathbf{y}$

In general, $\mathbf{y}$ and $\boldsymbol{\psi}$ are updated separately, due to different computational costs. Evaluating $\log p(\mathbf{y}, \boldsymbol{\psi}[, \mathbf{t}]|\boldsymbol{\xi})$ for different values of $\mathbf{y}$ is inexpensive, since only $\log p(\mathbf{t}|\mathbf{y})$ and $-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}$ need to be re-computed. In contrast, evaluating $\log p(\mathbf{y}, \boldsymbol{\psi}[, \mathbf{t}]|\boldsymbol{\xi})$ for different values of $\boldsymbol{\psi}$ requires re-computations of the inverse $\mathbf{C}^{-1}$ and the determinant $|\mathbf{C}|$, with computational complexity in the order of $O(N^3)$.

Assuming L samples $\boldsymbol{\theta}_1$, …, $\boldsymbol{\theta}_L$ have been drawn from $\log p(\mathbf{y}, \boldsymbol{\psi}|\mathbf{t}, \boldsymbol{\xi})$, the set of samples can be used to approximate the distribution of the latent $Y(\mathbf{x}_{N+1})$ given $\mathbf{t}$, given by

$$p(Y(\mathbf{x}_{N+1})|\mathbf{t}, \mathbf{x}_{N+1}, \boldsymbol{\xi}) = \int p(Y(\mathbf{x}_{N+1})|\mathbf{y}, \mathbf{x}_{N+1}, \boldsymbol{\psi})p(\mathbf{y}, \boldsymbol{\psi}|\mathbf{t}, \boldsymbol{\xi})d\mathbf{y}d\boldsymbol{\psi}$$
$$\approx \frac{1}{L}\sum_{l=1}^{L} p(Y(\mathbf{x}_{N+1})|\mathbf{y}_l, \boldsymbol{\psi}_l, \mathbf{t}, \mathbf{x}_{N+1})$$

Subsequently, the predictive distribution can be approximated, resulting in

$$p(T(\mathbf{x}_{N+1}) = c_k|\mathbf{t}, \mathbf{x}_{N+1}, \boldsymbol{\xi}))$$
$$\approx \frac{1}{L}\sum_{l=1}^{L} p(T(\mathbf{x}_{N+1}) = c_k|Y(\mathbf{x}_{N+1}))p(Y(\mathbf{x}_{N+1})|\mathbf{y}_l, \boldsymbol{\psi}_l, \mathbf{t}, \mathbf{x}_{N+1})$$

Chapter 5

Prediction for large data sets

Summary In chapter 3 and 4, Gaussian process techniques have been introduced, with Gaussian process techniques for classification developed from geostatistical methods and Gaussian process techniques for regression. In this chapter, results of research on the topic of extending the applicability of Gaussian process techniques to increasingly large data sets are reviewed, focusing on approaches based on obtaining a reduced-rank approximation to the covariance matrix and sparse approximation techniques applicable in Gaussian process classification.

Gaussian process models can be applied to predict quantities in the continuous case (where inference is exact) and the discrete case (where inference is approximate) as described in previous chapters. A problem with Gaussian process prediction is computational complexity (in the order of $O(N^3)$), resulting from inversion of a $N \times N$ covariance matrix $\mathbf{C}$. E.g., in case of regression,

$$p(T(\mathbf{x}_{N+1})|\mathbf{t}) = N(T(\mathbf{x}_{N+1})|\mathbf{k}(\mathbf{C} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{t}, k - \mathbf{k}(\mathbf{C} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{k}^T),$$

or, in case of classification,

$q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1}) = \Phi(\frac{\mathbf{c}\mathbf{C}^{-1}\hat{\mathbf{y}}}{\sqrt{(1+c-\mathbf{c}(\mathbf{C}+\mathbf{W}^{-1})^{-1}\mathbf{c}^T)}})$ in case of the Laplace approximation, and $q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1}) = \Phi(\frac{\mathbf{c}(\mathbf{C}+\tilde{\mathbf{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}},}{\sqrt{(1+c-\mathbf{c}(\mathbf{C}+\tilde{\mathbf{\Sigma}})^{-1}\mathbf{c}^T)}})$ in case of Expectation Propagation, respectively.

In general, this makes Gaussian process prediction impractical or infeasible for large data sets, with $N > 10000$ data points.

In order to overcome this limitation, a number of approximations has been suggested. In general, these techniques can be divided into approaches substituting the covariance matrix with a reduced rank approximation and sparse Gaussian process approaches based on the selection of a set of $M < N$ inducing inputs to represent the data, constituting an active set I. While the former can be applied to regression and classification, only few sparse GP techniques can be applied to classification tasks.

In the following sections, the two approaches are reviewed, with focus on classification in case of sparse Gaussian process techniques. In the first, algebraic techniques aiming at obtaining a reduced rank approximation to $\mathbf{C}$ are presented. In the following section, an unifying view of sparse approximations for Gaussian process techniques (due to (34)) is introduced, focusing on sparse Gaussian process techniques suitable for classification tasks.

5.1 Reduced rank approximations

In general, the optimal reduced-rank approximation $\tilde{\mathbf{C}}_Q$ (of rank Q) to $\mathbf{C}$ [with respect to the Frobenius norm [1] $\|\mathbf{C}\|_F = tr(\mathbf{C}\mathbf{C}^T)$] is $\mathbf{U}_Q\mathbf{\Lambda}_Q\mathbf{U}_Q^T$, where $\mathbf{\Lambda}_Q$ is the matrix of the leading Q eigenvalues of $\mathbf{C}$ and $\boldsymbol{U}_Q$ is the matrix of the corresponding eigenvectors. Unfortunately (due to a time complexity in the order of N^3), the eigenvalue decomposition offers little computational advantage.

Nyström approximation The Nyström approximation can be applied to obtain a reduced-rank approximation to the covariance matrix. The technique can be motivated from an eigenanalysis of the covariance function $c(\mathbf{x}_i, \mathbf{x}_j)$:

$\int c(\mathbf{x}_i, \mathbf{x}_j)\phi(\mathbf{x}_i)p(\mathbf{x}_i)d\mathbf{x_i} = \lambda\phi(\mathbf{x}_j)$

with the set of orthogonal eigenfunctions $\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots$ with the property $\int \phi_k(\mathbf{x})\phi_l(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \delta_{kl}$, corresponding eigenvalues $\lambda_1, \lambda_2, \ldots$ (assuming an ordering so that $\lambda_1 \geq \lambda_2 \geq \ldots$)

Based on above expression, an approximation to the eigenfunctions and eigenval-

[1] see e.g. (14)

ues can be obtained from:

$$\lambda_k \phi_k(\mathbf{x}_j) = \int c(\mathbf{x}_i, \mathbf{x}_j)\phi_k(\mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{1}{M}\sum_{m=1}^{M} c(\mathbf{x}_m, \mathbf{x}_j)\phi_k(\mathbf{x}_m)$$

In context of GP prediction, the Nyström method can be used to approximate the eigenvalues/ eigenvectors of $\mathbf{C}$. The procedure starts with a subset (the active set) I of $M < N$ data points (with $J = \{\mathbf{x}_i | \mathbf{x}_i \notin I\}$). Then, $\mathbf{C}$ can be partitioned as follows:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{I,I} & \mathbf{C}_{I,J} \\ \mathbf{C}_{J,I} & \mathbf{C}_{J,J} \end{pmatrix}$$

, with $\mathbf{C}_{A,B}$ denoting a covariance matrix with entries determined by the covariance function $c(\mathbf{x}_i, \mathbf{x}_j)$, with $\mathbf{x}_i \in A$, $\mathbf{x}_j \in B$.

Given the eigenvalues and eigenvectors of $\mathbf{C}_{I,I}$ (denoted $\lambda_k^{(M)}$ and $\mathbf{u}_k^{(M)}$, respectively), the eigenvalues/ eigenvectors of $\mathbf{C}$ can be approximated:

$$\tilde{\lambda}_k^{(N)} = \frac{N}{M}\lambda_k^{(M)}\ ^{1},\ k = 1, \ldots, M,$$

$$\tilde{\mathbf{u}}_k^{(N)} = \sqrt{\frac{M}{N}}\frac{1}{\lambda_k^{(M)}}\mathbf{C}_{NM}\mathbf{u}_k^{(M)},\ k = 1, \ldots, M,$$

with $\mathbf{C}_{NM} = (\mathbf{C}_{MN})^T = [\mathbf{C}_{I,I}, \mathbf{C}_{J,I}]^T$.

In general, the covariance matrix can be approximated up to rank Q, with $Q \leq M$. Choosing the first Q=M eigenvalues/ eigenvectors (according to the ordering $\tilde{\lambda}_1^{(Q)} \geq \tilde{\lambda}_2^{(Q)} \geq \ldots \geq \tilde{\lambda}_Q^{(Q)}$) results in

$$\mathbf{C} \approx \tilde{\mathbf{C}}_Q = \mathbf{C}_{NQ}\sum_{q=1}^{Q}\frac{1}{\lambda_q^{(Q)}}\mathbf{u}_q^{(Q)}(\mathbf{u}_q^{(Q)})^T\mathbf{C}_{QN}$$
$$= \mathbf{C}_{NM}\mathbf{C}_{MM}^{-1}\mathbf{C}_{MN},$$

with $\mathbf{C}_{MM} = \mathbf{C}_{I,I}$.

Having obtained $\tilde{\mathbf{C}}_Q$, the expression for the predictive distribution in case of regression can be written

[1] From $M\lambda_k\phi_k(\mathbf{x}_j) \approx \sum_{m=1}^{M} c(\mathbf{x}_m, \mathbf{x}_j)\phi_k(\mathbf{x}_m)$, for $j = 1, \ldots, M$

$$p(T(\mathbf{x}_{N+1})|\mathbf{t}) = N(T(\mathbf{x}_{N+1})|\tilde{\mathbf{k}}(\tilde{\mathbf{C}}_Q + \sigma^2\mathbf{I})^{-1}\mathbf{t}), \tilde{k} - \tilde{\mathbf{k}}(\tilde{\mathbf{C}}_Q + \sigma^2\mathbf{I})^{-1}\tilde{\mathbf{k}}^T)$$
$$= N(T(\mathbf{x}_{N+1})|\mathbf{c}(\mathbf{x}_{N+1})\mathbf{Q}^{-1}\mathbf{C}_{QN}\mathbf{t}, \sigma^2\mathbf{c}(\mathbf{x}_{N+1})\mathbf{Q}^{-1}\mathbf{c}(\mathbf{x}_{N+1})^T),$$

with $\tilde{\mathbf{k}} = (\tilde{k}(\mathbf{x}_{N+1}, \mathbf{x}_1), \ldots, \tilde{k}(\mathbf{x}_{N+1}, \mathbf{x}_N))$, $\tilde{k} = \tilde{k}(\mathbf{x}_{N+1}, \mathbf{x}_{N+1})$, where

$$\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \textstyle\sum_{q=1}^{Q} \frac{\lambda_q^{(Q)}}{Q} \frac{Q}{(\lambda_q^{(Q)})^2} \mathbf{c}^T(\mathbf{x}_i)\mathbf{u}_q^{(Q)}(\mathbf{u}_q^{(Q)})^T\mathbf{c}(\mathbf{x}_j)$$
$$= (c(\mathbf{x}_1, \mathbf{x}_i), \ldots, c(\mathbf{x}_Q, \mathbf{x}_i))^T\mathbf{C}_{QQ}^{-1}(c(\mathbf{x}_1, \mathbf{x}_j), \ldots, c(\mathbf{x}_Q, \mathbf{x}_j)), \text{ and}$$

$$(\tilde{\mathbf{C}}_Q + \sigma^2\mathbf{I})^{-1} = (\mathbf{C}_{NQ}\mathbf{C}_{QQ}^{-1}\mathbf{C}_{QN} + \sigma^2\mathbf{I})^{-1}$$
$$= \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{C}_{NQ}(\mathbf{C}_{QQ} + \mathbf{C}_{QN}\mathbf{C}_{NQ})^{-1}\mathbf{C}_{QN} = \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{C}_{NQ}\mathbf{Q}^{-1}\mathbf{C}_{QN} \ ,$$

with $\mathbf{Q} = (\mathbf{C}_{QQ}+\mathbf{C}_{QN}\mathbf{C}_{NQ})$, and application of the Sherman-Morrison-Woodbury identity (see Appendix B.2 for details) resulting in reduction of time complexity to $O(Q^3)$.

In a similar way, the predictive distribution in case of classification can be written

$q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1}) = \Phi(\frac{\tilde{\mathbf{k}}\tilde{\mathbf{C}}_Q^{-1}\hat{\mathbf{y}}}{\sqrt{(1+\tilde{k}-\tilde{\mathbf{k}}(\tilde{\mathbf{C}}_Q+\mathbf{W}^{-1})^{-1}\tilde{\mathbf{k}}^T)}})$ in case of the Laplace approximation, and $q(T(\mathbf{x}_{N+1}) = 1|\mathbf{t}, \mathbf{x}_{N+1}) = \Phi(\frac{\tilde{\mathbf{k}}(\tilde{\mathbf{C}}_Q+\tilde{\mathbf{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}},}{\sqrt{(1+\tilde{k}-\tilde{\mathbf{k}}(\tilde{\mathbf{C}}_Q+\tilde{\mathbf{\Sigma}})^{-1}\tilde{\mathbf{k}}^T)}})$ in case of Expectation Propagation.

Sparse Greedy Matrix Approximation The Nyström approximation was derived in (48) for application to kernel machines. An alternative view resulting in the same approximation is due to Smola and Schölkopf (42) For a data point $\mathbf{x}$ (with $\mathbf{x} \in J$) the technique is based on approximating the covariance $c(\mathbf{x}_i, \mathbf{x})$ by a linear combination of covariances $c(\mathbf{x}_j, \mathbf{x})$, with $j \in I$:

$$c(\mathbf{x}_i, \mathbf{x}) \approx \hat{c}(\mathbf{x}_i, \mathbf{x}) = \textstyle\sum_{j \in I} \tilde{c}_{ij} c(\mathbf{x}_j, \mathbf{x})$$

for $\tilde{c}_{ij} \in \mathbb{R}$, $i \in N \setminus I$, $j \in I$.

In order to obtain the $\tilde{c}_{ij}$, the expression

$$Err(\tilde{\mathbf{C}}) = \textstyle\sum_{i=1}^{N} \|c(\mathbf{x}_i, \mathbf{x}) - \hat{c}(\mathbf{x}_i, \mathbf{x})\|_{\mathcal{H}}^2 \ [1]$$

[1] with $\|\cdot\|_{\mathcal{H}}^2$ denoting proximity in the Reproducing Kernel Hilbert Space $\mathcal{H}$ induced by the covariance function $c(\cdot)$.

$$= tr(\mathbf{C}) - 2tr(\tilde{\mathbf{C}}\mathbf{C}_{MN}) + tr(\tilde{\mathbf{C}}\mathbf{C}_{MM}\tilde{\mathbf{C}}^T)$$

with respect to $\tilde{\mathbf{C}}$ is minimized, resulting in

$$\tilde{\mathbf{C}}_* = \arg\min_{\tilde{\mathbf{C}}} E(\tilde{\mathbf{C}}) = \mathbf{C}_{NM}\mathbf{C}_{MM}^{-1}$$

Smola and Schölkopf (42) suggest a greedy algorithm to choose points to include into the active set in a way that minimizes $E(\tilde{\mathbf{C}})$. Since it takes MN operations to evaluate the change in E due to the inclusion of a new data point, Smola and Schölkopf suggest finding the best point from a random subset of $\{\mathbf{x}_i\}$, $i \in N \setminus I$ on each iteration.

5.2 Sparse GP techniques

To overcome the computational limitations of Gaussian process prediction, several approximate inference techniques have been proposed, aiming at improving the scalability of Gaussian processes. In general, these methods, referred to as sparse approximations, share the property that only a subset of latent variables is treated exactly in inference, and the remaining variables are given approximate treatment. Quinonero-Candela et al. (34) provide a unifying view of sparse approximations for Gaussian processes, extending (33). In the following section, the common framework is presented, introducing the concepts of inducing variables (inducing inputs), inducing conditionals, and effective prior.

Inducing variables In order to introduce the framework, the GP prior over the latent $\mathbf{y}$ and $Y(\mathbf{x}_{N+1})$ $(\mathbf{y}_{N+})$ [1] is rewritten by introducing a set of additional M (with $M < N$) latent variables $(u_1, \ldots, u_M)$, referred to as inducing variables, corresponding to a set of inducing inputs $\mathbf{X}_u = (\mathbf{x}_{u_1}, \ldots, \mathbf{x}_{u_M})^T$:

$$p(\mathbf{y}_{N+}, \mathbf{y}) = \int p(\mathbf{y}_{N+}, \mathbf{y}, \mathbf{u})d\mathbf{u} = \int p(\mathbf{y}_{N+}), \mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u},$$

where $\mathbf{u} \sim N(\mathbf{u}|\mathbf{0}, \mathbf{C}_{\mathbf{u},\mathbf{u}})$ [2]

[1] with $\mathbf{y}_{N+}$ denoting a vector of latent values $(y_{N+1}, y_{N+2}, \ldots)$ at (test) data points $\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \ldots$.

[2] with $\mathbf{C}_{\mathbf{a},\mathbf{b}}$ denoting a covariance matrix with entries determined by the covariance function $c(\cdot)$, evaluated at data points corresponding to the latent variables $\mathbf{a}$, $\mathbf{b}$.

In the unifying framework, the inducing variables (inducing dependencies between $\mathbf{y}$ (the elements of the training set) and $\mathbf{y}_{N+}$ (the elements of the test set)) constitute the active set I. Particular sparse algorithms choose the inducing variables in different way. Some techniques choose the inducing inputs to be a subset of the training set (or test set), while others do not.

Inducing conditionals The second assumption in the common framework is the assumption of conditional independence of $\mathbf{y}$ and $\mathbf{y}_{N+}$, given $\mathbf{u}$:

$$p(\mathbf{y}_{N+}, \mathbf{y}) = \int p(\mathbf{y}_{N+}|\mathbf{u})p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \approx q(\mathbf{y}_{N+}, \mathbf{y})$$
$$= \int q(\mathbf{y}_{N+}|\mathbf{u})q(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u}$$

In the framework (34), different sparse approximation techniques correspond to different additional assumptions with respect to the inducing conditionals $q(\mathbf{y}|\mathbf{u})$ (training conditional) and $q(\mathbf{y}_{N+}|\mathbf{u})$ (test conditional), with

$$p(\mathbf{y}|\mathbf{u}) = N(\mathbf{y}|\mathbf{C}_{y,u}\mathbf{C}_{u,u}^{-1}\mathbf{u}, \mathbf{C}_{y,y} - \mathbf{Q}_{y,y}) \approx q(\mathbf{y}|\mathbf{u}),$$

$$p(\mathbf{y}_{N+}|\mathbf{u}) = N(\mathbf{y}_{N+}|\mathbf{C}_{\mathbf{y}_{N+},u}\mathbf{C}_{u,u}^{-1}\mathbf{u}, \mathbf{C}_{\mathbf{y}_{N+},\mathbf{y}_{N+}} - \mathbf{Q}_{\mathbf{y}_{N+},\mathbf{y}_{N+}}) \approx q(\mathbf{y}_{N+}|\mathbf{u}),$$

with $\mathbf{Q}_{a,b} = \mathbf{C}_{a,u}\mathbf{C}_{u,u}^{-1}\mathbf{C}_{u,b}$.

Particular choices of $q(\mathbf{y}|\mathbf{u})$ and $q(\mathbf{y}_{N+}|\mathbf{u})$ result in a particular form of the effective prior $q(\mathbf{y}, \mathbf{y}_{N+})$, given by

$$q(\mathbf{y}, \mathbf{y}_{N+}) = N(\mathbf{0}, \begin{pmatrix} cov(\mathbf{y}, \mathbf{y}) & cov(\mathbf{y}, \mathbf{y}_{N+}) \\ cov(\mathbf{y}_{N+}, \mathbf{y}) & cov(\mathbf{y}_{N+}, \mathbf{y}_{N+}) \end{pmatrix})$$[1]

Different sparse approximation techniques implement different selection criteria with respect to the choice of the inducing variables. The inducing variables can be chosen from a subset of training (data) points, following a greedy selection scheme (21). Alternatively, $\mathbf{u}$ can be chosen from the set of test (data) points (45). Finally, by relaxing the constraint that the inducing inputs must be a subset of training/ test points, the set of inducing variables can be obtained via optimization with respect to inducing inputs, as proposed in (43).

[1] with the shorthand $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Quinonero-Candela et al. (34) assign different techniques to the common framework. Specifically, they identify the following general approaches:

Subset of Data (SoD),
with training conditional $q(\mathbf{y}|\mathbf{u}) = p(\mathbf{y}|\mathbf{u})$, test conditional $q(\mathbf{y}_{N+}|\mathbf{u}) = p(\mathbf{y}_{N+}|\mathbf{u})$ and effective prior

$$q(\mathbf{y}, \mathbf{y}_{N+}) = N(\mathbf{0}, \begin{pmatrix} \mathbf{C}_{y,y} & \mathbf{C}_{y,\mathbf{y}_{N+}} \\ \mathbf{C}_{\mathbf{y}_{N+},y} & \mathbf{C}_{\mathbf{y}_{N+},\mathbf{y}_{N+}} \end{pmatrix},$$

Deterministic Inducing Conditional (DIC),
with $q(\mathbf{y}|\mathbf{u}) = N(\mathbf{y}|\mathbf{C}_{y,u}\mathbf{C}_{u,u}^{-1}\mathbf{u}, \mathbf{0})$, $q(\mathbf{y}_{N+}|\mathbf{u}) = N(\mathbf{y}_{N+}|\mathbf{C}_{\mathbf{y}_{N+},u}\mathbf{C}_{u,u}^{-1}\mathbf{u}, \mathbf{0})$ and

$$q(\mathbf{y}, \mathbf{y}_{N+}) = N(\mathbf{0}, \begin{pmatrix} \mathbf{Q}_{y,y} & \mathbf{Q}_{y,\mathbf{y}_{N+}} \\ \mathbf{Q}_{\mathbf{y}_{N+},y} & \mathbf{Q}_{\mathbf{y}_{N+},\mathbf{y}_{N+}} \end{pmatrix},$$

Deterministic Training Conditional (DTC),
with $q(\mathbf{y}|\mathbf{u}) = N(\mathbf{y}|\mathbf{C}_{y,u}\mathbf{C}_{u,u}^{-1}\mathbf{u}, \mathbf{0})$, $q(\mathbf{y}_{N+}|\mathbf{u}) = p(\mathbf{y}_{N+}|\mathbf{u})$ and

$$q(\mathbf{y}, \mathbf{y}_{N+}) = N(\mathbf{0}, \begin{pmatrix} \mathbf{Q}_{y,y} & \mathbf{Q}_{y,\mathbf{y}_{N+}} \\ \mathbf{Q}_{\mathbf{y}_{N+},y} & \mathbf{C}_{\mathbf{y}_{N+},\mathbf{y}_{N+}} \end{pmatrix},$$

Fully Independent (Training) Conditional (FI(T)C),
with $q(\mathbf{y}|\mathbf{u}) = N(\mathbf{y}|\mathbf{C}_{y,u}\mathbf{C}_{u,u}^{-1}\mathbf{u}, diag(\mathbf{C}_{y,y} - \mathbf{Q}_{y,y}))$ [1], $q(\mathbf{y}_{N+}|\mathbf{u}) = p(\mathbf{y}_{N+}|\mathbf{u})$ and

$$q(\mathbf{y}, \mathbf{y}_{N+}) = N(\mathbf{0}, \begin{pmatrix} \mathbf{Q}_{y,y} - diag(\mathbf{Q}_{y,y} - \mathbf{C}_{y,y}) & \mathbf{Q}_{y,\mathbf{y}_{N+}} \\ \mathbf{Q}_{\mathbf{y}_{N+},y} & \mathbf{C}_{\mathbf{y}_{N+},\mathbf{y}_{N+}} \end{pmatrix}, \text{ and}$$

Partially Independent Training Conditional (PI(T)C),
with $q(\mathbf{y}|\mathbf{u}) = N(\mathbf{y}|\mathbf{C}_{y,u}\mathbf{C}_{u,u}^{-1}\mathbf{u}, blockdiag(\mathbf{C}_{y,y} - \mathbf{Q}_{y,y}))$ [2], $q(\mathbf{y}_{N+}|\mathbf{u}) = p(\mathbf{y}_{N+}|\mathbf{u})$ and

$$q(\mathbf{y}, \mathbf{y}_{N+}) = N(\mathbf{0}, \begin{pmatrix} \mathbf{Q}_{y,y} - blockdiag(\mathbf{Q}_{y,y} - \mathbf{C}_{y,y}) & \mathbf{Q}_{y,\mathbf{y}_{N+}} \\ \mathbf{Q}_{\mathbf{y}_{N+},y} & \mathbf{C}_{\mathbf{y}_{N+},\mathbf{y}_{N+}} \end{pmatrix}$$

Quinonero-Candela et al. give the computational complexity for each approach. Specifically, time complexity is in the order of $O(M^3)$ (training), $O(M)$ (predic-

[1] with $\mathbf{D} = diag(\mathbf{A})$ denoting a $N \times N$ diagonal matrix $\mathbf{D}$ with entries $d_{ii} = a_{ii}$, $i = 1, \ldots, N$.

[2] with $\mathbf{B} = blockdiag(\mathbf{A})$ denoting a $N \times N$ block diagonal matrix $\mathbf{B}$, with block entries matching the entries of $\mathbf{A}$.

tion) and (M^2) (uncertainty estimates) in case of Subset of Data (SoD), with training set size M, and in the order of $O(NM^2)$ (training), $O(M)$ (prediction) and $O(M^2)$ (uncertainty estimates) for the DTC, DIC, FI(T)C, and PI(T)C approximations, with training set size N.

Sparse GP for classification In analogy to the non-sparse case, sparse GP methods for classification have to deal with the problem of non-Gaussian likelihood, introducing the need for approximations, using methods described in the previous chapter. Consequently, only a subset of the techniques applicable in case of Gaussian process regression can be applied for classification.

Sparse GP techniques for classification have been proposed in (21), (5), (6), and (45). In (21), Lawrence et al. propose a Subset of Data (SoD) approximation implementing a greedy forward selection scheme according to the differential entropy score $\Delta_i = H(Q(\mathbf{y}_{I\cup\{i\}})) - H(Q(\mathbf{y}_I))$ (with $Q(\mathbf{y}_I)$ denoting the approximate posterior $q((\mathbf{y}_j)_{j\in I}|(t_j)_{j\in I})$, and $H(Q(\mathbf{y}_I))$ denoting the entropy of $Q(\mathbf{y}_I)$) for candidate data points $\mathbf{x}_i$, $i \in N \setminus I$.[1] After inclusion of a data point, an approximation to $q((\mathbf{y}_j)_{j\in I}|(t_j)_{j\in I})$ is made using the EP algorithm. In order to reduce time complexity, Lawrence et al. (21) update $\mathbf{\Sigma} = (\mathbf{C}^{-1} + \tilde{\mathbf{\Sigma}}^{-1})^{-1}$ (the covariance matrix of the approximate Gaussian posterior) by sequentially growing the Cholesky factor $\mathbf{L}$, with $\mathbf{L}\mathbf{L}^T = \mathbf{I} + \tilde{\mathbf{\Sigma}}_I^{\frac{1}{2}} \mathbf{C}_I \tilde{\mathbf{\Sigma}}_I^{\frac{1}{2}}$[2], so that $\mathbf{\Sigma} = \mathbf{C} - \mathbf{C}_I \tilde{\mathbf{\Sigma}}_I^{\frac{1}{2}} (\mathbf{L}\mathbf{L}^T)^{-1} \tilde{\mathbf{\Sigma}}_I^{\frac{1}{2}} \mathbf{C}_I$ (avoiding explicit inversion of $(\mathbf{C}^{-1} + \tilde{\mathbf{\Sigma}}^{-1})$). In (41), the differential entropy score was replaced by the information gain criterion $\Delta_j = D(Q(\mathbf{y}_{I\cup\{j\}})||Q(\mathbf{y}_I))$ (with $D(q(\mathbf{y})||p(\mathbf{y}))$ denoting the Kullback-Leibler divergence between $q(\mathbf{y})$ and $p(\mathbf{y})$), including a term dependent on $(t_j)_{j\in I}$. The resulting algorithm, referred to as the Informative Vector Machine (IVM), was extended to apply a variational approximation to the marginal likelihood for hyperparameter estimation ((40), (41)).

In an alternative approach, Csato and Opper (5) propose a Deterministic Training Conditional (DTC) approximation, based on sequential construction of a (non-Gaussian) posterior $p_{z+1}(\mathbf{y}|\mathbf{t})$, given by $p_{z+1}(\mathbf{y}|\mathbf{t}) = \frac{p(t_{z+1}|\mathbf{y})q_z(\mathbf{y})}{\int p(t_{z+1}|\mathbf{y})q_z(\mathbf{y})d\mathbf{y}}$, which is

[1] This corresponds to the selection of the data point resulting in the greatest reduction in the variance of the approximate marginal posterior $q(\mathbf{y}_{N+}|\mathbf{t})$.

[2] with $\tilde{\mathbf{\Sigma}}_I$, $\mathbf{C}_I$ denoting the respective covariance matrices restricted to the set I

projected to the closest Gaussian posterior $q_{z+1}(\mathbf{y}|\mathbf{t})$. [1] In the approach, sparsity is introduced by expressing the covariance $c(\mathbf{x}_{z+1}, \mathbf{x})$ (i.e., the covariance for the data point considered in the $z+1$-th iteration) in terms of covariances $c(\mathbf{x}_1, \mathbf{x}), \ldots, c(\mathbf{x}_z, \mathbf{x})$ (in a way similar to (42)):

$$c(\mathbf{x}_{z+1}, \mathbf{x}) \approx \hat{e}_{z+1}^T \mathbf{C}_{1,\ldots,z;1,\ldots,N}$$
$$= (c(\mathbf{x}_1, \mathbf{x}_{z+1}), \ldots, c(\mathbf{x}_z, \mathbf{x}_{z+1}))\mathbf{C}^{-1}_{1,\ldots,z;1,\ldots,z}\mathbf{C}_{1,\ldots,z;1,\ldots,N} = \mathbf{u}^T\mathbf{C}^{-1}_{u,u}\mathbf{C}_{u,y},$$

with $\hat{e}_{z+1} = \mathbf{C}^{-1}_{1,\ldots,z;1,\ldots,z}(c(\mathbf{x}_1, \mathbf{x}_{z+1}), \ldots, c(\mathbf{x}_z, \mathbf{x}_{z+1}))^T$, so that

$$q_{SOGP}(\mathbf{y}|\mathbf{u}) = N(\mathbf{y}|\mathbf{C}_{y,u}\mathbf{C}^{-1}_{u,u}\mathbf{u}, \mathbf{0}).$$

In the approach of (Csato and Opper 2002), the covariance $c(\mathbf{x}_{z+1}, \mathbf{x})$ is approximated by $(c(\mathbf{x}_1, \mathbf{x}_{z+1}), \ldots, c(\mathbf{x}_z, \mathbf{x}_{z+1}))\mathbf{C}^{-1}_{1,\ldots,z;1,\ldots,z}\mathbf{C}_{1,\ldots,z;1,\ldots,N}$ if the approximation error introduced by the projection of $c(\mathbf{x}_i, \mathbf{x})$ to the subspace spanned by $(c(\mathbf{x}_1, \mathbf{x}), \ldots, c(\mathbf{x}_z, \mathbf{x}))^T$, given by $|q_{z+1}|(c(\mathbf{x}_{z+1}, \mathbf{x}_{z+1}) - \mathbf{c}_{z+1}\mathbf{C}_z^{-1}\mathbf{c}_{z+1})$, with $q_{z+1} = \frac{\partial \log p(t_{z+1}|y_{z+1})}{\partial E(t_{z+1})}$ and $\mathbf{c}_{z+1} = (c(\mathbf{x}_{z+1}, \mathbf{x}_1), \ldots, c(\mathbf{x}_{z+1}, \mathbf{x}_z))$, does not exceed a threshold ϵ. In order to improve the quality of the approximation, data points can be removed (pruned) from the index set (of size M), in exchange for other data points.

Finally, the Bayesian committee machine (BCM) (45), introduced as a sparse GP technique for regression, has been generalized to deal with classification tasks. For the BCM, the data is split in K subsets (denoted D_k, for $k = 1, \ldots, K$), with $D_k = (\mathbf{X}_k, \mathbf{t}_k) = ((\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,N_k}), (t_{k,1}, \ldots, t_{k,N_k}))$. Assuming that $p(\mathbf{t}_k|\mathbf{t}_{k-1}, \mathbf{y}_{N+}) \approx p(\mathbf{t}_k|\mathbf{y}_{N+})$, it holds that

$$p(\mathbf{y}_{N+}|\mathbf{t}) \propto \frac{\prod_{k=1}^K p(\mathbf{y}_{N+}|\mathbf{t}_k)}{p(\mathbf{y}_{N+})^{K-1}}$$

Subsequently, Laplace's approximation is applied to each of the K subsets to yield the approximate predictive mean $E(\mathbf{y}_{N+}|\mathbf{t}_k)$ and covariance $cov(\mathbf{y}_{N+}|\mathbf{t}_k)$.

Due to the assumption $p(\mathbf{t}_k|\mathbf{t}_{k-1}, \mathbf{y}_{N+}) \approx p(\mathbf{t}_k|\mathbf{y}_{N+})$, the generalized BCM has been interpreted as an instance of the PITC approximation, with inducing points

[1] As in the case of Expectation Propagation, the optimal projection results from matching the (first and second) moments of $p_{z+1}(\mathbf{y}|\mathbf{t})$ and $q_{z+1}(\mathbf{y}|\mathbf{t})$.

given by (test) data points $\mathbf{t}_{N+}$. The computational complexity of the approximation depends on the structure of the (block-diagonal) covariance $blockdiag(\mathbf{Q}_{y,y} - \mathbf{C}_{y,y})$. For $K = N/M$ blocks of size $M \times M$ each, the time complexity is in the order $O(NM^2)$.

Chapter 6

Application to spatial data

In previous chapters, Gaussian process techniques have been introduced, focusing on regression (chapter 3) and classification (chapter 4) techniques, considering the special case of large data sets (chapter 5). In this chapter, Gaussian process classification techniques are applied to two data sets describing the occurence of different types of mass movements, including earth movements and snow avalanches. In general, the occurence of these movements can be traced to a combination of specific factors (morphological, meteorological, anthropogenic, or other) contributing to disposition (tendency) of rock, earth, or snow to move downslope, or triggering mass movement. The particular combination of factors resulting in movement depends on the type of the natural phenomenon and is typically specific and local.

In the following sections, application of Gaussian process classification techniques to two real-world data sets is described, starting with the first, describing the occurence of earth movements in the administrative region Hochtannberg (Vorarlberg, Austria). Subsequently, GP classification techniques are applied to a data set describing the occurence of snow avalanches in the region Lochaber (Scotland, UK), subject to numerous avalanche events during winter season.

6.1 Susceptibility to earth movements

In this section, application of Gaussian process classification techniques to a data set describing the occurence of earth movements in the region Hochtannberg (Vorarlberg, Austria), with the aim of probabilistic classification of susceptibility on regional scale and probabilistic mapping is described. The description of the data set follows (10) , (11) , (12) , summarizing work focusing on a comparison of predictive performance of probabilistic classification techniques on the task (10) and implementation of Gaussian process classification techniques in a way allowing for flexible use in context of a suitable early warning chain (11), (12) and is given in the first subsection. In the following subsection, the choice of techniques from (10) is extended by including Gaussian process classification techniques described in chapter 4 (Expectation Propagation) and chapter 5 (the Sparse Online Gaussian Process classifier (5)). These techniques are evaluated on the task and results are summarized.

6.1.1 Study area

Based on work carried out in context of the project 'Georisk map Vorarlberg' by the Dept. of Applied Geology, University of Karlsruhe (TH), the study area Hochtannberg was chosen as basis for a data set describing the occurence of earth movements. The study area is a region of 115 km^2 size, located in Vorarlberg, the westernmost federal state of Austria, with border to Germany to the north, and Liechtenstein and Switzerland to the west. The geography, climate, and geology of study area are described in detail in (37). Throughout the area, Hochtannberg is considered susceptible to different types of mass movement (ibid.). The mass movements occuring in the area are of interest because of the high density of touristic infrastructure built at steep slopes.

6.1.2 Data set

6.1.2.1 Data set/ preprocessing

The data set consisting of a set of thematic layers describing different dispositive factors was produced using GIS technology (ESRI ArcInfo 9.2). A digital elevation model (DEM) and topographic and geologic maps covering the area of Vorarlberg were provided by the Land Surveying Office Vorarlberg (Vorarlberg, Austria). The digital elevation model (with resolution of 5 m) was used to calculate morphometric features, describing the morphology of the study area (including slope, curvature, and slope aspect). Digital data on geology (lithology and tectonics) was extracted from a geologic map of Vorarlberg, resulting in a grid of 25 m grid point size. Additionally, a grid of Euclidean distances to tectonic faults was calculated, with shortest distance to a tectonic fault assigned to each grid point. Digital land cover data for Vorarlberg, supplied by Umweltbundesamt GmbH, was derived from satellite data (Landsat 5). Available land cover data for Vorarlberg consists of 12 land cover types, including built-up area, agricultural area, forests and natural area, wetlands and water surfaces. Digital data on rainfall intensity, describing the number of days with precipitation exceeding 10 mm averaged over a period of 30 years (1970-2000), was made available by (36).

In order to apply classification techniques to study area, a data set was built drawing on results from the project 'Georisk map Vorarlberg'. In context of the project, occurences of earth movements were mapped, resulting in a layer containing locations of occurences. For pre-processing, all data layers (including feature layers and slide inventory) were converted to ESRI ASCII grid format, with size corresponding to the size of the study area, and resolution of 25 m/ grid point.

In course of pre-processing, a subset of features (including morphometric features obtained from the digital elevation model and thematic maps (elevation, flow accumulation, distance to faults)) was standardized. A second set of features (including categorical and non-linear features (lithology, land cover, precipitation, slope aspect) was encoded as a set of binary features (in case of categorical features), or as a set of intervals on the range of values taken by the original feature.

6.1.3 Methods

In order to evaluate the performance of the Gaussian process techniques on the task, three different methods were applied to the data set, including a Gaussian process classifier based on Laplace's approximation, a classifier based on Expectation Propagation, and a classifier based on (5), referred to as the Sparse Online Gaussian Process (SOGP) classifier. All techniques were implemented in the statistical computing environment R[1], with parts of the code written in C. In addition to the Gaussian process classifiers, a probabilistic variant of the Support Vector Machine[2] ((3) due to ((31) made available in kernlab (Karatzoglou et al. 2004) was applied to the task. In case of Gaussian process techniques, classification was performed using an anisotropic squared exponential covariance (kernel) $c_{SE}(\mathbf{x}_i, \mathbf{x}_j) = \theta_0 \exp(-\frac{1}{2}\boldsymbol{\theta}^T(\mathbf{x}_j - \mathbf{x}_i)^2)$, with covariance pa-

[1] http://www.r-project.com

[2] C-SVM

rameter $\boldsymbol{\theta}$ initialized to to the mean of parameter values on the training set. For the SVM, an isotropic squared exponential kernel function was used (with $c_{SE}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2\sigma^2}(\mathbf{x}_j - \mathbf{x}_i)^2)$, in order to find optimal values for parameters σ and C by search on a balanced validation set, with size $N_{val} = 480$.

6.1.4 Results

Classification performance Results from the different classifiers obtained on the data set are summarized in Table 6.1. For each classifier, the model was trained on a balanced data set of $N = 2400$ data points, represented by feature vectors of length 41 and a (binary) class label t_i, indicating if data point $\mathbf{x}_i$ was part of an area where movement occurence was registered ($t_i = 1$), or not ($t_i = 0$) (with negative examples sampled uniformly from a set of $(176277 - 1200)$ grid points, after including all grid points with $t_i = 1$), subject to k-fold ($k = 5$) cross-validation. In order to determine the influence of training set size on classification performance, training sets of decreasing size were used, corresponding to fractions of N. In case of the SOGP classifier, active sets of different (maximum) sizes were used, corresponding to fractions of N.

The figures in Table 6.1 indicate that the Gaussian process classifiers based on Laplace's approximation (LA-GP), Expectation Propagation (EP-GP) and the SOGP classifier (SOGP-GP) yield results comparable to the SVM (with θ and C found on the validation set) in case of $N = 1920$, with decreasing classification performance for training sets of decreasing size in case of LA-GP and EP-GP. Training times[1] for Gaussian process classifiers based on Expectation Propagation and Laplace's approximation range between 2 sec. ($N = 480$) and 20 sec. ($N = 1920$) (Laplace's approximation) and 20 sec. ($N = 480$) and 20 min. ($N = 1920$) (Expectation Propagation). Training times for the SOGP classifier range between 2 min. (for an active set of maximum size $M = 480$) and 3 min. (for $M = 1920$). The reduced computational complexity of the SVM training algorithm (SMO ((30))) results in training times ranging between 1 min. ($N = 480$) and 6 min. ($N = 1920$) (including time required for tuning of θ and C). Finally, a Gaussian process classifier based on Hybrid Monte Carlo was applied to the task, with application restricted to $N = 480$ data points due to the high computational cost.

[1] Intel T 2500 2 GHz,2 GB RAM

Data set size (grid points)	1920	1440	960	480
Method				
Laplace	87.4	86.3	84.6	81.7
EP	89.6	87.5	85.8	83.5
SOGP	88.4	87.9	87.6	86.5
SVM	92.9	91.8	90.9	88.9
HMC(*)	-	-	-	83.8

Table 6.1: Classification results (in % of correctly classified data points) for different classification techniques on different training set sizes for the Hochtannberg data.

Probabilistic mapping After training, the Gaussian process techniques and the SVM were used to generate susceptibility maps for the study area. For probabilistic mapping, the complete data set ($N = 2400$ grid points) was used, and trained models were applied to the study area, containing 176449 grid points. Resulting susceptibility maps were generated as 8 bit grey scale images, with the value of each pixel given by the (approximate) predictive probability $q(T(\mathbf{x}_i) = 1|\mathbf{t})$. For the Gaussian process classifiers, uncertainty maps were generated, following a similar procedure (substituting $var(q(Y(\mathbf{x}_{N+1})|\mathbf{t}))$ for the predictive probability for $q(T(\mathbf{x}_i) = 1|\mathbf{t})$). For further use/ post-processing, each map was exported in a number of formats, including PNG, TIFF, and ESRI ASCII Grid (using methods contained in the rgdal R package).

6.1.5 Discussion

From a comparison of classification performance and training times for Gaussian process classifiers and the SVM, it can be observed that application of Gaussian process classifiers results in classification performance comparable to the SVM, at slightly longer training times in case of Gaussian process classifiers based on Laplace's approximation and Expectation Propagation, and slightly shorter training times in case of the Sparse Online Gaussian Process (SOGP) classifier. From this observation, application of the SOGP classifier is considered the preferred option. For smaller active sets (with $M < N$), application of the SOGP classifier is advantageous, reducing prediction time complexity (required for the computation of the variance of $q(Y(\mathbf{x}_{N+1})|\mathbf{t})$) from $O(N^2)$ to $O(M^2)$ for each data point.

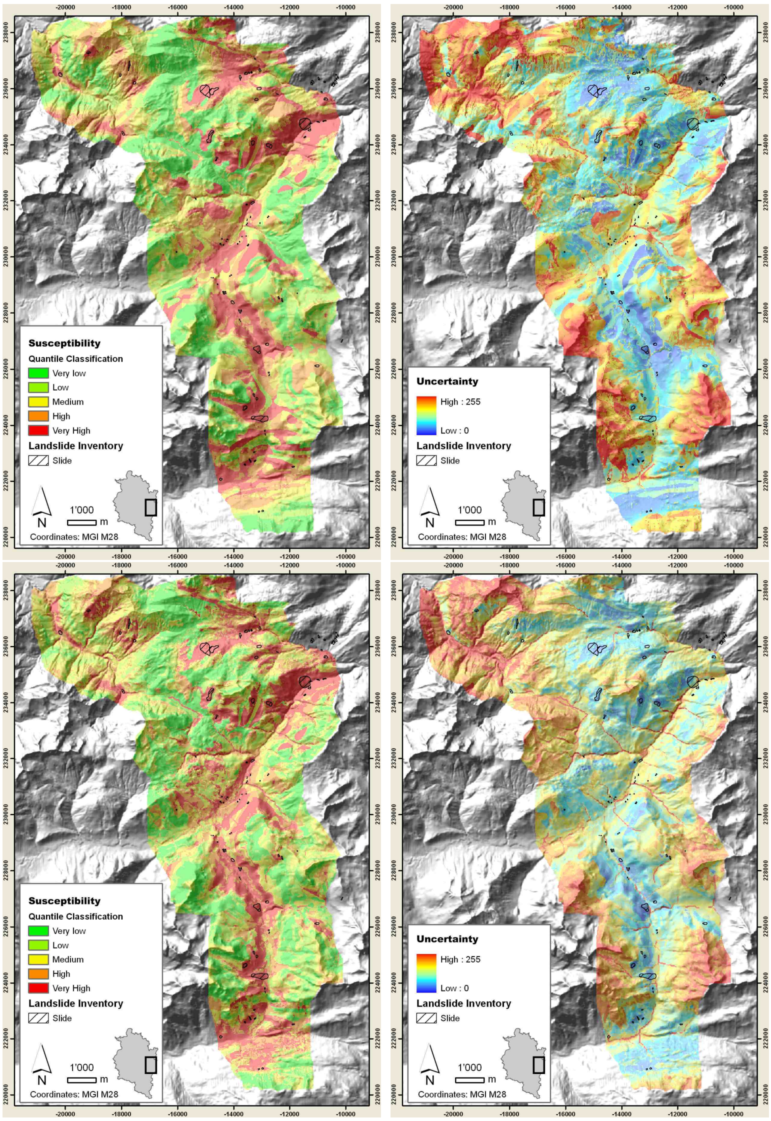

Figure 6.1: **Susceptibility/ uncertainty maps for study area Hochtannberg (1)**
Results of probabilistic classification for study area Hochtannberg. The GP classifier based on Laplace's approximation ((a), (b)) yields results comparable to the GP classifier based on Expectation Propagation ((c), (d)), with color gradient indicating probability of movement occurence ((a), (c)) and uncertainty in prediction ((b), (d)).
a) result of probabilistic classification (Laplace's approximation/ susceptibility map); b) result of probabilistic classification (Laplace's approximation/ uncertainty map); c) result of probabilistic classification (Expectation Propagation/ susceptibility map); d) result of probabilistic classification (Expectation Propagation/ uncertainty map).

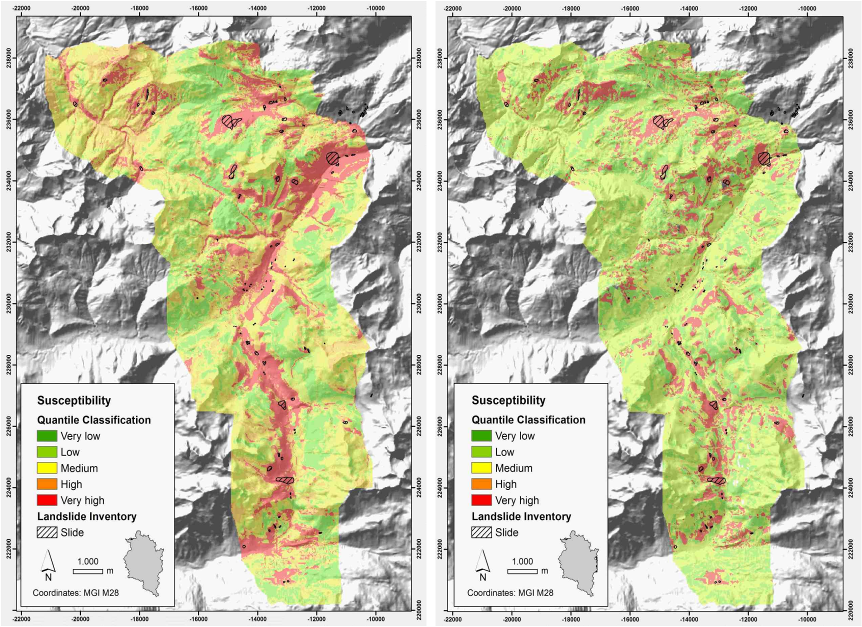

Figure 6.2: **Susceptibility maps for study area Hochtannberg (2)**
Results of probabilistic classification for study area Hochtannberg. The SOGP classifier of (Csato and Opper 2002) (a) yields predictive performance comparable to the performance of the (optimally tuned) SVM (b), with color gradient indicating probability of movement occurence. Additionally, the GP classifier provides information related to uncertainty in prediction (Fig. 6.3 (b)).
a) SOGP, $M = 480$, susceptibility; b) SVM, susceptibility.

In practice, reduction of prediction time is crucial, in particular in case when the (test) set (the set of data points (locations) to be considered in prediction) is large, which is the case in probabilistic mapping.

In terms of prediction time, computation of uncertainty in prediction (the variance of $q(Y(\mathbf{x}_{N+1})|\mathbf{t}))$ makes application of Gaussian process classifiers more computationally demanding. The problem of prediction time complexity is addressed by sparse GP classification techniques (e.g., the SOGP classifier). In case of the SVM, the problem is circumvented by not considering uncertainty in prediction (resulting in shorter prediction times). However, uncertainty estimates obtained from the predictive disctribution provide additional information, which is of interest when predictions are made when real-world data is considered (where observations may be missing). A particular example for such a case is shown in Fig. 6.3.

6.2 Avalanche hazard

In the second case study, Gaussian process classification techniques were applied to a data set describing the occurence of snow avalanches in the Lochaber region, Scotland, a well-known ski venue for which daily avalanche forecast is available. In contrast to the first case study (with focus on classification of susceptibility and probabilistic mapping), the prediction problem in the second case study is an instance of a spatio-temporal prediction problem, with data points describing environmental conditions at a set of avalanche path sites, including weather conditions measured by local forecasters or registered by an automatic weather station. Given a set of examples with data points and observations describing occurences of events on days of the winter season (4 months per year) in previous years, the prediction task is an accurate forecast of avalanche event occurences at the level of avalanche paths, focusing on reliable forecast (sensitive prediction) of avalanche events.

6.2.1 Data set

The description of the data follows (23). For each data point in the data set, corresponding to one of 40 avalanche path sites on one of 1131 days of the winter seasons between 1991 and 2007, a set of features derived from measurements of weather and snowpack conditions (precipitation, snow drift, air temperature,

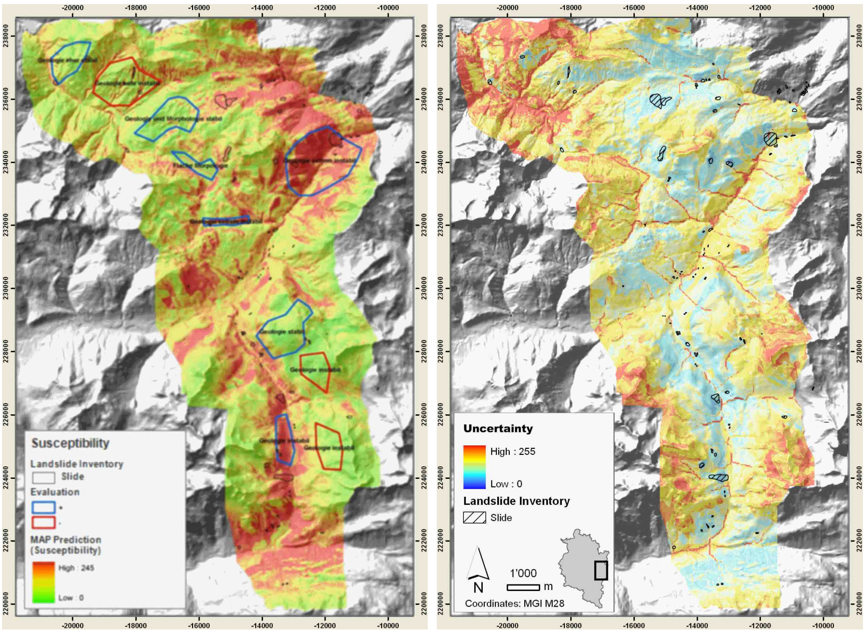

Figure 6.3: **Susceptibility/ uncertainty maps for study area Hochtannberg (3)**
Results of probabilistic classification for study area Hochtannberg. a) result of probabilistic classification (Laplace's approximation) subject to qualitative evaluation ((36)), with blue polygons indicating agreement, red polygons indicating disagreement with the model. The two susceptible areas (red polygons) in the south east of the study area are misclassified by both the GP classifier(s) and the SVM (Fig. 6.1, Fig. 6.2 (a), (b)). In the example, the two polygons indicating disagreement are related to medium to high prediction uncertainty (in (b)).

wind speed and direction, cloud cover, foot penetration, and snow temperature) was combined with a morphological description derived from a Digital Elevation Model (DEM), with a resolution of 10 m. In the resulting data set, a binary target value indicating whether an avalanche event was registered (or not) was assigned to each data point (one of 40 avalanche path sites on one of 1131 days), with the description consisting of a set of 39 variables, including 22 features corresponding to local morphological and meteorological conditions at each path site and 17 temporal features, describing weather and snowpack conditions for each data point.

Based on this data, three data sets were built, including a balanced training set consisting of 894 data points and observations from the period between 1991 and 2005, a validation set consisting of 10277 data points and observations from the same period (including 148 observations of avalanche events registered at one of the path sites) and 10129 negative examples, and an (unbalanced) test set consisting of 4792 data points and observations (including 72 positive examples and 4720 negative examples) from the period between 2006 and 2007. In this arrangement, the validation set was introduced for the purpose of model selection for Gaussian process classifiers and SVM, allowing for optimization with respect to different performance criteria, including a measure of sensitivity referred to as the Hansen-Kuipers discriminant (HK), defined as

$$HK = \frac{|TP|}{|TP|+|FN|} - \frac{|FP|}{|FP|+|TN|} \in [-1;1]$$

, with TP denoting the set of true positives (with $t_i = 1 = sign(q(T(\mathbf{x}_i)|\mathbf{t}) - 0.5)$ for $\mathbf{x}_i \in TP$), FP denoting the set of false positives, or false alarms (with $t_i = 1 \neq sign(q(T(\mathbf{x}_i)|\mathbf{t}) - 0.5)$ for $\mathbf{x}_i \in FP$), TN denoting the set of true negatives (with $t_i = 0 = sign(q(T(\mathbf{x}_i)|\mathbf{t}) - 0.5)$ for $\mathbf{x}_i \in TN$), and FN denoting the set of false negatives (with $t_i = 0 \neq sign(q(T(\mathbf{x}_i)|\mathbf{t}) - 0.5)$ for $\mathbf{x}_i \in FN$).

6.2.2 Methods

In order to evaluate the performance of Gaussian process techniques on the task of reliable avalanche forecast (sensitive prediction), three different Gaussian process methods (LA-GP, EP-GP, SOGP-GP) were applied to the data set, following a training procedure involving model fitting (training) on the training set, with model selection on the validation set. For model selection, two different approaches were chosen, including hyperparameter estimation via optimization of the marginal likelihood of a subset of the validation set[1] (using an isotropic squared exponential covariance $c_{SE}(\mathbf{x}_i, \mathbf{x}_j) = \theta_0 \exp(-\frac{1}{2}\theta(\mathbf{x}_j - \mathbf{x}_i)^2))$ and grid search on the validation set (using an isotropic squared exponential covariance (kernel)), aiming at maximization of the HK score.

6.2.3 Results

Classification performance Results for the Gaussian process techniques (LA-GP, EP-GP, SOGP-GP) and the SVM on the prediction task are summarized in Table 6.2, with predictive performance and sensitivity indicated by (1 - misclassi-

[1] with $N = 1036$

fication error) (ME) and HK score, with parameters determined by optimization of the marginal likelihood (denoted LA-GP-LH, EP-GP-LH), and grid search, involving searching for θ_{0opt} and θ_{opt} on the set $\{4^s|s = -5,\ldots,5\}$, followed by search on $\{4+0.8t|t = -4,\ldots,5\}$ for θ_0 and $\{2^{-(t+4)}|t = 1,\ldots,6\}$ for θ, resulting in $\theta_{0opt} = 8$, $\theta_{opt} = 2^{-4} = 0.0625$ (LA-GP-GS), $\theta_{0opt} = 6.4$, $\theta_{opt} = 2^{-4} = 0.0625$ (EP-GP-GS) and $\theta_{0opt} = 4$, $\theta_{opt} = 2^{-4} = 0.0625$ (SOGP-GP-GS). In order to assess the performance of GP classifiers compared to the SVM, an implementation of the SVM (e1071 (8)) was applied to the data, using a training set consisting of 30837 data points. For the SVM, parameters θ and C were set to $\theta_{opt} = \frac{1}{2.4^2}$ and $C_{opt} = 5.6$, resulting from search on the set $\{4^s|s = -5,\ldots,5\}$, followed by search on the set $\{4 + 0.8t|t = -4,\ldots,5\}$ for $\sigma_{opt} = \sqrt{\frac{1}{\theta_{opt}}}$ and C_{opt}. Having obtained these parameters, training times for Gaussian process techniques ranged between 15 sec. (LA-GP-LH, LA-GP-GS), 3 min. (SOGP-GP-GS) and 27 sec. (EP-GP-LH, EP-GP-GS), with longer training time (27 min.) for the SVM.

Results in Table 6.2 indicate that Gaussian process classifiers based on Laplace's approximation (LA-GP) and Expectation Propagation (EP-GP) and the Sparse Online Gaussian Process classifier (SOGP-GP) (with different active set sizes corresponding to fractions (percentage) of training set size) yield results comparable to the SVM, with comparable number of true positives and false positives (false alarms) in case of model selection based on grid search (LA-GP-GS, EP-GP-GS, SOGP-GP-GS), and HK scores indicating comparable sensitivity on part of Gaussian process techniques. Results of Gaussian process classification with (hyper-)parameters determined by optimization of the marginal likelihood (LA-GP-LH, EP-GP-LH) indicate lower performance on the task, as indicated by $|TP|$, $|FP|$, and HK.

Probabilistic mapping After training, the Gaussian process classifiers and the SVM were applied to a data grid consisting of $500^2 = 250000$ data points, describing weather and snowpack conditions in the study area on a particular day in winter season 2007 (14.02.2007), combined with a morphological description derived from a Digital Elevation Model (DEM) (with resolution 10 m/ grid point). Similar to the Hochtannberg case study, avalanche hazard maps resulting from predictions made under the Gaussian process models and the SVM (Fig. 6.4, Fig. 6.5, Fig. 6.6) were generated as 8-bit grey scale images, with the value of

Performance criterion	ME	HK	$\lvert TP\rvert$	$\lvert FP\rvert$
Method				
LA-GP-LH	74.2	0.546	58	1222
LA-GP-GS	82.6	0.714	64	825
EP-GP-LH	74.9	0.553	58	1189
EP-GP-GS	82.9	0.718	64	808
SOGP-GP-GS-25	78.9	0.674	63	998
SOGP-GP-GS-50	80.2	0.692	64	939
SOGP-GP-GS-75	80.4	0.693	64	929
SOGP-GP-GS-100	80.5	0.693	64	922
SVM	79.2	0.665	63	990

Table 6.2: Classification results (in % of correctly classified data points (1-ME), the HK discriminant score (HK), the number of true positives ($|TP|$) and the number of false positives ($|FP|$)) on the Lochaber data set (test set, $N = 4792$).

each pixel given by the (approximate) predictive probability $q(T(\mathbf{x}_i) = 1|\mathbf{t})$.

6.2.4 Discussion

The figures in Table 6.2 suggest that application of Gaussian process classifiers results in classification performance comparable to the SVM, with Hansen-Kuipers discriminant scores resulting from model selection on the validation set suggesting comparable predictive performance and sensitivity at forecasting tasks. A comparison of probabilistic maps resulting from predictions of the Gaussian process classifiers and the SVM (Fig. 6.4, Fig. 6.5, Fig. 6.6) indicates that the Gaussian process classifiers (LA-GP-GS, EP-GP-GS) yield accurate forecasts for the 14.02.2007, with assignment of probability above decision threshold ($p(T(\mathbf{x}) = 1) > .5$) to five out of five grid points (four out of five grid points in case of EP-GP-GS) where avalanche occurence was registered (Fig. 6.5, Fig. 6.6), a result comparable to the result of prediction obtained from the SVM (Fig. 6.4). An inspection of hazard maps in Fig 6.4-6.6 shows that results of the SVM indicate a more pronounced discrimination, with probabilities $p(T(\mathbf{x}) = 1)$ in the range $[0.03; 0.99]$, in contrast to $q(T(\mathbf{x}) = 1) \in [0.36; 0.73]$ for (LA-GP-GS), $q(T(\mathbf{x}) = 1) \in [0.24; 0.90]$ for (EP-GP-GS), and $q(T(\mathbf{x}) = 1) \in [0.40; 0.74]$ for (SOGP-GP-GS). Results of Gaussian process classifiers indicate more conservative classification, with uncertainty estimates indicating lower uncertainty (higher confidence) in prediction at grid points where avalanche occurence was registered (Fig. 6.7-6.9 (b)).

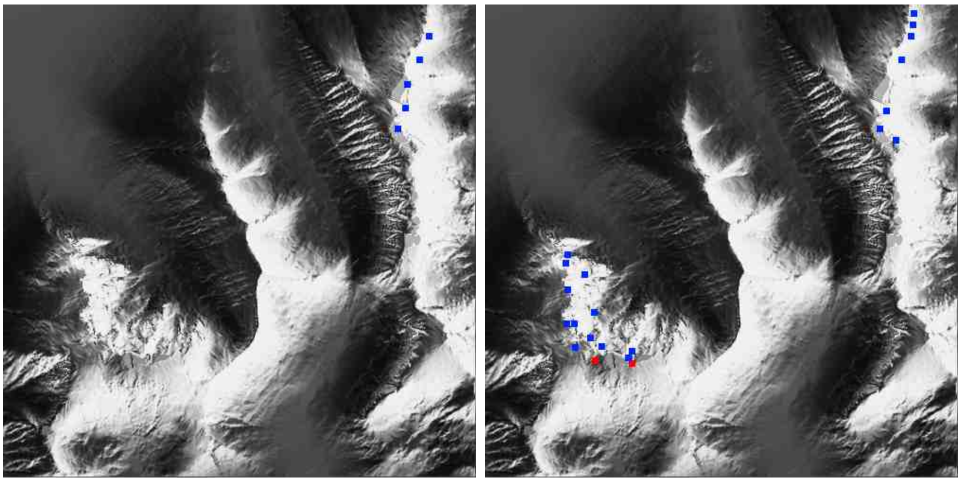

Figure 6.4: **Avalanche hazard maps for study area Lochaber (1)**
Results of probabilistic avalanche forecast for the Lochaber study area (14.02.2007, SVM). a) The SVM classifier assigns high probability of avalanche occurence ($p(T(\mathbf{x}) = 1) > .72$) to 5 out of 5 grid points corresponding to locations where avalanche occurence was registered (blue points, with brighter shades of grey indicating higher probability). b) Results of avalanche forecast for 14.02.2007 assign probabilities above decision threshold ($p(T(\mathbf{x}) = 1) > .5$) to 40 out of 48 grid points corresponding to locations where avalanche occurence was registered prior to 14.02.2007 (16.01.2006-28.01.2007) (blue/ red points).

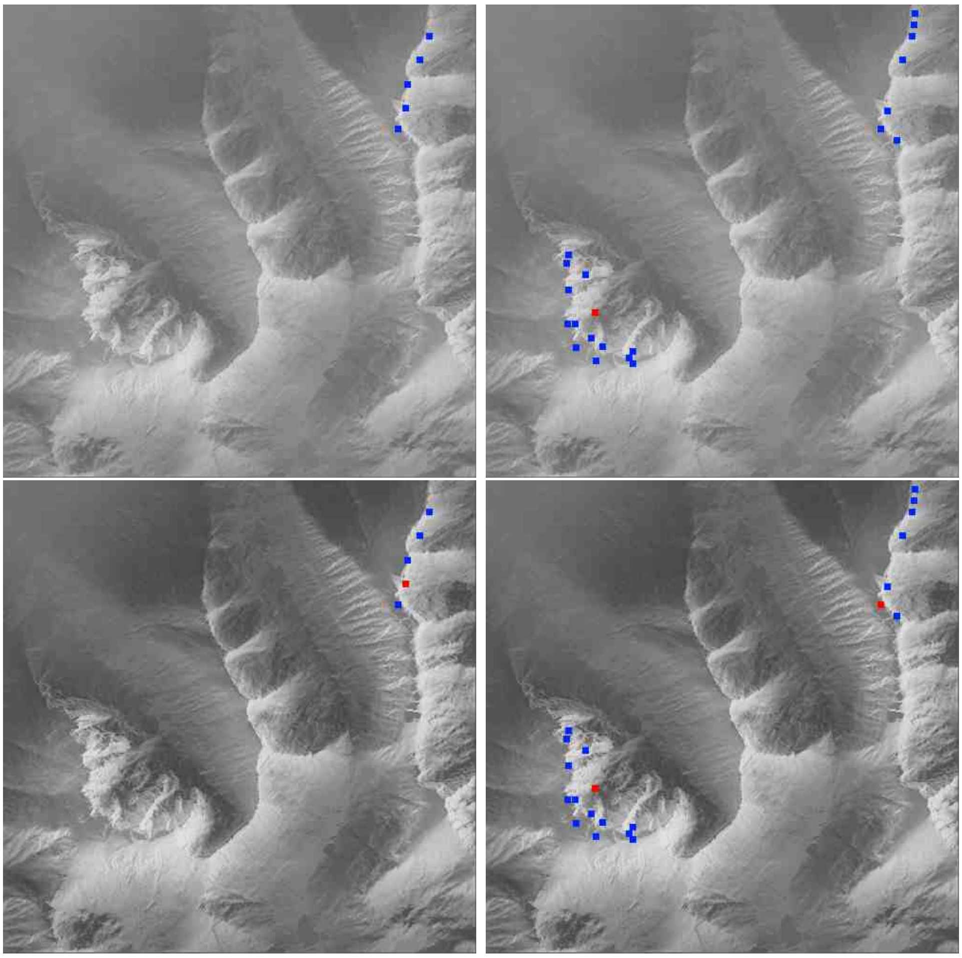

Figure 6.5: **Avalanche hazard maps for study area Lochaber (2)**
Results of probabilistic avalanche forecast for the Lochaber study area (14.02.2007, Gaussian processes). For the GP classifier based on Laplace's approximation, probabilities $q(T(\mathbf{x}) = 1) > .5$ are assigned to 5 out of 5 grid points corresponding to locations where avalanche occurence was registered on 14.02.2007 (a). For the GP classifier based on Expectation Propagation, probabilities $q(T(\mathbf{x}) = 1) > .5$ are assigned to 4 out of 5 grid points ((c), blue/ red points, with brighter shades of grey indicating higher probability).
a) result of avalanche forecast for 14.02.2007 (Laplace's approximation/ hazard map); b) Results of avalanche forecast for the 14.02.2007 assign probabilities above decision threshold ($q(T(\mathbf{x}) = 1) > .5$) to 45 out of 48 grid points corresponding to locations where avalanche occurence was registered prior to 14.02.2007 (16.01.2006-28.01.2007) (blue/ red points). c) result of avalanche forecast for 14.02.2007 (Expectation Propagation/ hazard map); d) Results of avalanche forecast for the 14.02.2007 assign probabilities above decision threshold ($q(T(\mathbf{x}) = 1) > .5$) to 42 out of 48 grid points (c.f. (b)).

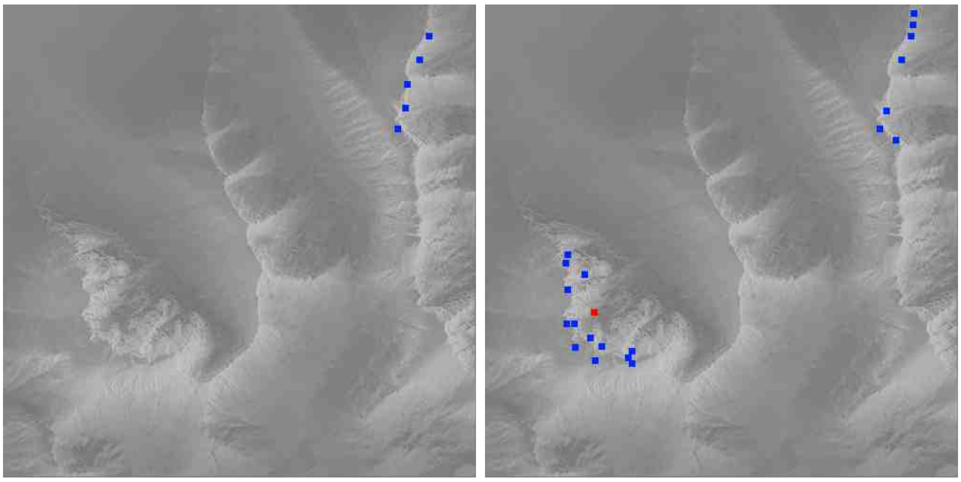

Figure 6.6: **Avalanche hazard maps for study area Lochaber (3)**
Results of probabilistic avalanche forecast for the Lochaber study area (14.02.2007, SOGP). a) The Sparse Online Gaussian Process (SOGP) classifier assigns probability above decision threshold ($q(T(\mathbf{x}) = 1) > .5$) to 5 out of 5 grid points corresponding to locations where avalanche occurence was registered (blue points, with brighter shades of grey indicating higher probability). b) Results of avalanche forecast for the 14.02.2007 assign probabilities above decision threshold ($q(T(\mathbf{x}) = 1) > .5$) to 48 out of 48 grid points corresponding to locations where avalanche occurence was registered prior to 14.02.2007 (16.01.2006-28.01.2007) (blue/ red points).

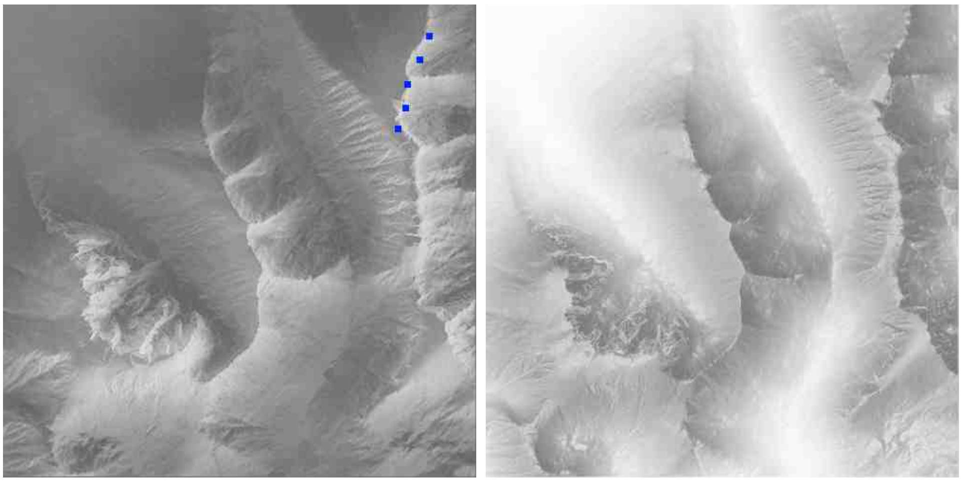

Figure 6.7: **Avalanche hazard/ uncertainty maps for study area Lochaber (4)**
Results of probabilistic avalanche forecast for the Lochaber study area (14.02.2007, Laplace's approximation). The uncertainty map (based on $var(q(Y(\mathbf{x}_{N+1})|\mathbf{t})))$ in (b) indicates lower uncertainty (higher confidence) in prediction at grid points where avalanche occurence was registered in the training set (with dark shades of grey indicating low uncertainty). Conversely, higher uncertainty is assigned to grid points corresponding to locations where no avalanche occurence was registered. a) result of avalanche forecast for 14.02.2007 (Laplace's approximation/ hazard map), with blue points corresponding to locations where avalanche occurence was registered; b) result of avalanche forecast for 14.02.2007 (Laplace's approximation/ uncertainty map).

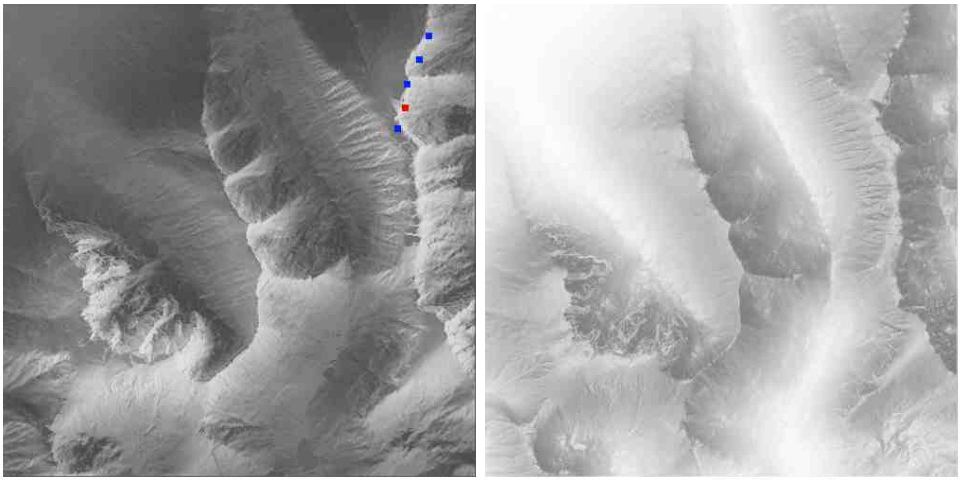

Figure 6.8: **Avalanche hazard/ uncertainty maps for study area Lochaber (5)**
Results of probabilistic avalanche forecast for the Lochaber study area (14.02.2007, EP). a) result of avalanche forecast for 14.02.2007 (Expectation Propagation/ hazard map), with blue points corresponding to locations where avalanche occurence was registered; b) result of avalanche forecast for 14.02.2007 (Exectation Propagation/ uncertainty map).

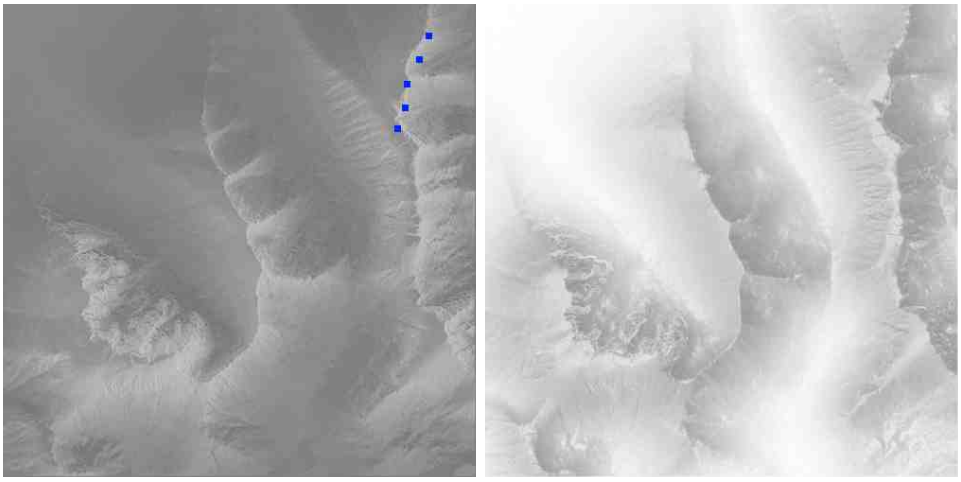

Figure 6.9: **Avalanche hazard/ uncertainty maps for study area Lochaber (6)**
Results of probabilistic avalanche forecast for the Lochaber study area (14.02.2007, SOGP). a) result of avalanche forecast for 14.02.2007 (SOGP/ hazard map), with blue points corresponding to locations where avalanche occurence was registered; b) result of avalanche forecast for 14.02.2007 (SOGP/ uncertainty map).

Chapter 7

Conclusions

In this work, the applicability of a discriminative probabilistic techniques (Gaussian process techniques) not previously applied in spatial prediction was investigated, focusing on the occurence of mass movements. The work summarizes results of research in context of research project 'Development of suitable information systems for early warning systems' (EGIFF)[1], with focus on introduction of techniques aiming at improvements in processing of data (measurements/ observations) in context of a suitable early warning chain. Results of this research include results of an application of different techniques for Gaussian process classification (involving application of stochastic and deterministic techniques for approximate inference) to two data sets describing the occurence of different types of mass movements (earth movements and snow avalanches). These results suggest applicability of Gaussian process techniques to classification of spatial/ spatio-temporal data on regional scale, with novelty resulting from application of different techniques for Gaussian process classification to high-dimensional real-world spatial/ spatio-temporal data sets. This is demonstrated by qualitative and quantitative evaluation, indicating predictive performance (sensitivity) comparable to the predictive performance (sensitivity) of the Support Vector Machine (SVM). Additionally, uncertainty estimates provided by Gaussian process classifiers result in additional information, which is of interest when predictions are made based on real-world data, where observations may be missing. An example for such a case is given in Fig. 6.3 (b) of chapter 6, with uncertainty estimates provided by Gaussian process classifiers indicating possible misclassification on part of both Gaussian process classifiers and the SVM, as suggested by qualitative

[1] part of the GEOTECHNOLOGIEN research program (http://www.geotechnologien.de), funded by the German Federal Ministry of Research and Education (BMBF)

evaluation (Fig. 6.3 (a)). Based on this result, availability of uncertainty estimates is considered advantageous, with justification in the assumption of missing observations (positive examples) due to difficult environmental conditions.

In order to apply Gaussian process techniques to classification tasks, a set of Gaussian process classification techniques making use of stochastic (Hybrid Monte Carlo) and deterministic (Laplace's Approximation, Expectation Propagation, SOGP) techniques for approximate inference was implemented in the statistical computing environment R[1], with critical parts of the code written in C. Resulting classification procedures can be applied to arbitrary classification tasks, involving spatial, spatio-temporal and non-spatial data. In context of the project, classification procedures can be applied in context of a suitable early warning chain, with prototypical client-server implementation (Java) available for platform-independent integration of classification procedures in existing information systems.

Gaussian process techniques for classification have rarely been applied to high-dimensional spatial classification problems. This work is one of the first that studied the applicability of several of these techniques to real-world spatial data. In this work, these techniques have been applied to classification tasks focusing on the occurence of mass movements (earth movements, snow avalanches). It is the author's hope to contribute to further study (more comprehensive evaluation) of these techniques.

[1] http://www.r-project.com

Appendices

Appendix A

The Gaussian

For $\mathbf{x} \in \mathrm{I\!R}^N$, the multivariate Gaussian distribution (denoted $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$) is parametrized by a mean vector $\boldsymbol{\mu} \in \mathrm{I\!R}^N$ and a (symmetric, positive definite) covariance matrix $\boldsymbol{\Sigma} \in \mathrm{I\!R}^{N\times N}$.

The multivariate Gaussian distribution has the form

$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\boldsymbol{\pi})^{\frac{N}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$

Conditioning and marginalizing Given a marginal Gaussian distribution p($\mathbf{x_1}$) and a conditional Gaussian distribution p($\mathbf{x_2}|\mathbf{x_1}$) in the form

$p(\mathbf{x_1}) = N(\mathbf{x_1}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, and $p(\mathbf{x_2}|\mathbf{x_1}) = N(\mathbf{x_2}|\mathbf{A}\mathbf{x_1} + \mathbf{b}, \mathbf{L}^{-1})$,

the marginal distribution $p(\mathbf{x_2})$ and the conditional distribution $p(\mathbf{x_1}|\mathbf{x_2})$ are

$p(\mathbf{x_2}) = N(\mathbf{x_2}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T)$, and $p(\mathbf{x_1}|\mathbf{x_2}) = N(\mathbf{x_1}|\boldsymbol{\Sigma}(\mathbf{A}^T\mathbf{L}(\mathbf{x_2} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}), \boldsymbol{\Sigma})$

, where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$.

Given a joint Gaussian distribution $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$, with

$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{11} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix}$

the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is given by

$p(\mathbf{x}_1|\mathbf{x}_2) = N(\mathbf{x}_1|\boldsymbol{\mu}_{\mathbf{1|2}}, \boldsymbol{\Sigma}_{1|2})$, where

$E(\mathbf{x}_1|\mathbf{x}_2) = \boldsymbol{\mu}_{\mathbf{1|2}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$, and $var(\mathbf{x}_1|\mathbf{x}_2) = \boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$

and the marginal distribution $p(\mathbf{x}_1)$ is given by

$p(\mathbf{x}_1) = N(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$

(see (46), sec. 9.3)

Product of Gaussians The product of two Gaussian distributions is an (unnormalized) Gaussian distribution:

$N(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)N(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = Z^{-1}N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

where $\boldsymbol{\mu} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_1\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2\boldsymbol{\mu}_2)$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}$, and

$Z^{-1} = (2\pi)^{-\frac{N}{2}}|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$.

Appendix B

Matrix results

B.1 Partitioned matrices

B.1.1

The matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & 0 \end{pmatrix}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ symmetric and non-singular, and $\mathbf{B} \in \mathbb{R}^{N\times P}$ and of full rank $P \leq N$, has inverse

$$\begin{pmatrix} \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{B}(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1} \\ (\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^{-1} & -(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{M}_{11}^- & \mathbf{M}_{12}^- \\ \mathbf{M}_{21}^- & \mathbf{M}_{22}^- \end{pmatrix}$$

B.1.2

The matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{a} \\ \mathbf{a}^T & 0 \end{pmatrix}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ symmetric and non-singular, and $\mathbf{a} \in \mathbb{R}^{N\times 1} \neq \mathbf{0}$, has inverse

$$\begin{pmatrix} \mathbf{A}^{-1}(\mathbf{I}_N - \mathbf{a}\frac{\mathbf{a}^T\mathbf{A}^{-1}}{\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}}) & \frac{\mathbf{A}^{-1}\mathbf{a}}{\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}} \\ \frac{\mathbf{a}^T\mathbf{A}^{-1}}{\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}} & \frac{-1}{\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}} \end{pmatrix}$$

B.2 Matrix identities

A useful matrix identity involving matrix inverses which can be used to derive several results is the following

$$(\mathbf{P}^{-1} + \mathbf{B}^T\mathbf{R}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{R}^{-1} = \mathbf{P}\mathbf{B}^T(\mathbf{B}\mathbf{P}\mathbf{B}^T + \mathbf{R})^{-1} \tag{B.1}$$

Another useful identity , known as the Sherman-Morrison-Woodbury identity (see e.g. (32)), is

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{B.2}$$

B.3 Matrix derivatives

The derivatives of the elements $\boldsymbol{\theta}$ of an inverse matrix $\mathbf{C}^{-1}$ are given by

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{C}^{-1} = -\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\boldsymbol{\theta}}\mathbf{C}^{-1} \tag{B.3}$$

The derivatives of the elements $\boldsymbol{\theta}$ of $\log|\mathbf{C}|$ are given by

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log|\mathbf{C}| = tr(\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\boldsymbol{\theta}}) \tag{B.4}$$

(see e.g. (15))

Appendix C

Derivation of kriging

To derive the kriging predictor in terms of the variogram function, start with the expression for the MSE:

$$MSE = E((\tilde{t}(\mathbf{x}_{N+1}) - T(\mathbf{x}_{N+1}))^2) = E((\lambda^T \mathbf{t} - T(\mathbf{x}_{N+1}))^2)$$

Noting that $E((T(\mathbf{x}_i) - T(\mathbf{x}_j))^2) = 2\gamma(\mathbf{x}_i - \mathbf{x}_j)$ (from the definition of the variogram function), the MSE is:

$$\begin{aligned}&MSE = E(-\tfrac{1}{2}\textstyle\sum_{i=1}^N \sum_{j=1}^N \lambda_i\lambda_j(T(\mathbf{x}_i) - T(\mathbf{x}_j))^2 + \sum_{i=1}^N \lambda_i(T(\mathbf{x}_i) - T(\mathbf{x}_{N+1}))^2)\\ &= -\textstyle\sum_{i=1}^N \sum_{j=1}^N \lambda_i\lambda_j\gamma(\mathbf{x}_i - \mathbf{x}_j) + 2\sum_{i=1}^N \lambda_i\gamma(\mathbf{x}_i - \mathbf{x}_{N+1})\\ &= -\boldsymbol{\lambda}^T\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^T\boldsymbol{\gamma}\end{aligned}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)^T \in \mathbb{R}^N$, $\boldsymbol{\gamma} = (\gamma(\mathbf{x}_1 - \mathbf{x}_{N+1}), \ldots, \gamma(\mathbf{x}_N - \mathbf{x}_{N+1}))^T \in \mathbb{R}^N$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$ is a $N\times N$ matrix with (i,j)-th element determined by $\gamma(\mathbf{x}_i - \mathbf{x}_j)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, N$.

Given a continuous, (conditionally) negative definite variogram model $\gamma(\mathbf{h})$, the expression for the MSE can be minimized subject to the unbiasedness constraint $\boldsymbol{\lambda}^T\mathbf{1} = 1$ in case of ordinary kriging (1) and $\boldsymbol{\lambda}^T\mathbf{F} = \mathbf{f}_{N+1}^T$ in case of universal kriging (2).

Introducing a scalar Lagrange multiplier α (1), and a $P \times 1$ Lagrange multiplier $\boldsymbol{\alpha}$ (2), the expressions to be minimized are

$-\boldsymbol{\lambda}^T\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^T\boldsymbol{\gamma} - 2\alpha(\boldsymbol{\lambda}^T\mathbf{1} - 1)$, and

$-\boldsymbol{\lambda}^T\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^T\boldsymbol{\gamma} - 2(\boldsymbol{\lambda}^T\mathbf{F} - \mathbf{f}_{N+1}^T)\boldsymbol{\alpha}$, respectively.

Differentiating with respect to $\boldsymbol{\lambda}$ and α (1), and $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ (2) and equating to 0 yields the matrix forms

$$\begin{pmatrix} \boldsymbol{\Gamma} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \alpha \end{pmatrix} = \begin{pmatrix} \gamma \\ 1 \end{pmatrix}, \text{ and } \begin{pmatrix} \boldsymbol{\Gamma} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} \gamma \\ \mathbf{f}_{N+1} \end{pmatrix}$$

Assuming that $\boldsymbol{\Gamma}$ is non-singular (this is the case when the variogram function is (conditionally) negative definite), and using the result for the inverse of a partitioned matrix $\mathbf{M}$ (see Appendix B for details), the solution (1) can be written

$$\begin{pmatrix} \boldsymbol{\lambda} \\ \alpha \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}^{-1}(\mathbf{I}_N - \mathbf{1}\frac{\mathbf{1}^T\boldsymbol{\Gamma}^{-1}}{\mathbf{1}^T\boldsymbol{\Gamma}^{-1}\mathbf{1}}) & \frac{\boldsymbol{\Gamma}^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Gamma}^{-1}\mathbf{1}} \\ \frac{\mathbf{1}^T\boldsymbol{\Gamma}^{-1}}{\mathbf{1}^T\boldsymbol{\Gamma}^{-1}\mathbf{1}} & \frac{-1}{\mathbf{1}^T\boldsymbol{\Gamma}^{-1}\mathbf{1}} \end{pmatrix} \begin{pmatrix} \gamma \\ 1 \end{pmatrix}$$

Hence, the solution has kriging weights

$$\boldsymbol{\lambda}_* = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma} + \mathbf{1}(\mathbf{1}^T\boldsymbol{\Gamma}^{-1}\mathbf{1})^{-1}(1 - \mathbf{1}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}))$$

in case of ordinary kriging, and, by analogy, $\boldsymbol{\lambda}_* = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma} + \mathbf{F}(\mathbf{F}^T\boldsymbol{\Gamma}^{-1}\mathbf{F})^{-1}(\mathbf{f}_{N+1} - \mathbf{F}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}))$

in case of universal kriging.

Substituting the kriging weights $\boldsymbol{\lambda}$ into the linear predictor $\tilde{t}(\mathbf{x}_{N+1}) = \lambda^T\mathbf{t}$ yields the kriging predictor, or BLUP.

Appendix D

IWLS

In general, the iterative weighted least squares algorithm to obtain a maximum likelihood solution for an unknown parameter $\boldsymbol{\theta}$ takes the form

$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - \mathbf{H}(\boldsymbol{\theta})^{-1} \frac{\partial l_{ML}}{\partial \boldsymbol{\theta}} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}$ in case of Newton-Raphson, or

$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - E(\mathbf{H}(\boldsymbol{\theta}))^{-1} \frac{\partial l_{ML}}{\partial \boldsymbol{\theta}} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}} = \boldsymbol{\theta}^{(m)} + \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial l_{ML}}{\partial \boldsymbol{\theta}} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}$ in the Fisher Scoring variant,

starting with an initial estimate $\boldsymbol{\theta}^{(0)}$,

where (m) indicates the mth iteration, $\mathbf{H}(\boldsymbol{\theta})$ denotes the Hessian matrix of second order derivatives of the log likelihood l_{ML}, and $\mathcal{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix, defined as $\mathcal{I}(\boldsymbol{\theta}) = -E(\mathbf{H}(\boldsymbol{\theta}))$.

In context of the GLM, an estimate for the unknown $\boldsymbol{\beta}$ is obtained through either variant, given the log likelihood for the model:

$l_{ML} = \log L_{ML} = \sum_{i=1}^{N} (t_i \gamma_i - b(\gamma_i))/\tau - \sum_{i=1}^{N} c(t_i, \tau)$

In order to evaluate the IWLS update step, it is necessary to obtain the Hessian $\mathbf{H}(\boldsymbol{\beta})$ or the Fisher information matrix $\mathcal{I}(\boldsymbol{\beta})$ by differentiating the log likelihood with respect to $\boldsymbol{\beta}$:

$\frac{\partial l_{ML}}{\partial \boldsymbol{\beta}} = \frac{\partial \log L_{ML}}{\partial \boldsymbol{\beta}}$

$$= \frac{1}{\tau^2}\sum_{i=1}^{N}(t_i\frac{\partial\gamma_i}{\partial\boldsymbol{\beta}} - \frac{\partial b(\gamma_i)}{\partial\gamma_i}\frac{\partial\gamma_i}{\partial\boldsymbol{\beta}}) \overset{\mu=\frac{\partial \mathrm{b}(\gamma)}{\partial\gamma}}{=} \frac{1}{\tau^2}\sum_{i=1}^{N}(t_i-\mu_i)\frac{\partial\gamma_i}{\partial\boldsymbol{\beta}} = \frac{1}{\tau^2}\sum_{i=1}^{N}(t_i-\mu_i)\frac{\partial\gamma_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}$$
$$= \frac{1}{\tau^2}\sum_{i=1}^{N}(t_i-\mu_i)\frac{\partial\gamma_i}{\partial\mu_i}(\frac{\partial h(\eta_i)}{\partial\eta_i})\frac{\partial\eta_i}{\partial\boldsymbol{\beta}} = \frac{1}{\tau^2}\sum_{i=1}^{N}(t_i-\mu_i)\frac{\partial\gamma_i}{\partial\mu_i}(\frac{\partial g(\mu_i)}{\partial\mu_i})^{-1}\mathbf{f}_i$$
$$= \frac{1}{\tau^2}\sum_{i=1}^{N}\frac{(t_i-\mu_i)}{v(\mu_i)(\frac{\partial g(\mu_i)}{\partial\mu_i})}\mathbf{f}_i = \frac{1}{\tau^2}\sum_{i=1}^{N}(t_i-\mu_i)w_i(\frac{\partial g(\mu_i)}{\partial\mu_i})\mathbf{f}_i,$$

after applying the chain rule to obtain $\frac{\partial\gamma_i}{\partial\boldsymbol{\beta}} = \frac{\partial\gamma_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}$, making use of the identities $\frac{\partial\gamma_i}{\partial\mu_i} = (\frac{\partial\mu_i}{\partial\gamma_i})^{-1} = (\frac{\partial^2 b(\gamma_i)}{\partial\gamma_i^2})^{-1} = \frac{1}{v(\mu_i)}$, and defining $w_i = \frac{1}{v(\mu_i)\frac{\partial^2 g(\mu_i)}{\partial\mu_i^2}}$.

In matrix form, the expression can be written

$$\frac{\partial l_{ML}}{\partial\boldsymbol{\beta}} = \frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{t}-\boldsymbol{\mu})$$

with $\mathbf{W} = diag(w_i)$ and $\boldsymbol{\Delta} = diag(\frac{\partial g(\mu_i)}{\partial\mu_i})$,

where $\mathbf{W}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\mu}$ involve the unknown $\boldsymbol{\beta}$.

Having obtained an expression for $\frac{\partial l_{ML}}{\partial\boldsymbol{\beta}}$, the expression for the Hessian can be derived:

$$\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial^2 l_{ML}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}(-1)\frac{\partial\mu}{\partial\boldsymbol{\beta}^T} + \frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\boldsymbol{\beta}^T}(\mathbf{t}-\boldsymbol{\mu})$$
$$\overset{\sum_{i=1}^{N}(\frac{\partial g(\mu_i)}{\partial\mu_i})^{-1}\mathbf{f}_i=diag((\frac{\partial g(\mu_i)}{\partial\mu_i})^{-1})\mathbf{F}}{=} -\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1}\mathbf{F} + \frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\boldsymbol{\beta}^T}(\mathbf{t}-\boldsymbol{\mu})$$
$$= -\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\mathbf{F} + \frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\boldsymbol{\beta}^T}(\mathbf{t}-\boldsymbol{\mu})$$

Hence, the Fisher information matrix is

$$\mathcal{I}(\boldsymbol{\beta}) = -E(\mathbf{H}(\boldsymbol{\beta})) = \frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\mathbf{F} + \frac{1}{\tau^2}\mathbf{F}^T\frac{\partial(\mathbf{W}\boldsymbol{\Delta})}{\partial\boldsymbol{\beta}^T}E(\mathbf{t}-\boldsymbol{\mu}) \overset{\mathrm{E}(\mathbf{t})=\boldsymbol{\mu}}{=} \frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\mathbf{F} + \mathbf{0}$$
$$= \frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\mathbf{F}$$

Using above expression, the Fisher Scoring update step for $\boldsymbol{\beta}$ can be written

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \tau(\mathbf{F}^T\mathbf{W}\mathbf{F})^{-1}\frac{1}{\tau^2}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{t}-\boldsymbol{\mu})$$
$$= \boldsymbol{\beta}^{(m)} + (\mathbf{F}^T\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{W}\boldsymbol{\Delta}(\mathbf{t}-\boldsymbol{\mu})$$

, or, equivalently,

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{F}^T\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{W}(\mathbf{F}\boldsymbol{\beta}^{(m)} + \boldsymbol{\Delta}(\mathbf{t} - \boldsymbol{\mu})).$$

with the Newton-Raphson scheme obtained by replacing $\mathcal{I}(\boldsymbol{\beta})$ by $-\mathbf{H}(\boldsymbol{\beta})$.

Bibliography

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover, 1964. 19, 39

[2] C. M. Bishop. *Neural Networks for Pattern Recognition* . Oxford University Press, 1995. 15

[3] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20, 1995. 67

[4] N. A. C. Cressie. *Statistics for spatial data*. Wiley, 1993. 9, 14

[5] L. Csato and M. Opper. Sparse On-Line Gaussian Processes. In *Neural Computation*, volume 14, 2002. 62, 66, 67

[6] L. Csato, M. Opper, and O. Winther. TAP Gibbs Free Energy, Belief Propagation and Sparsity. In *Advances in Neural Information Processing Systems*, volume 14, 2000. 62

[7] P. Diggle and P. J. Ribeiro. *Model-based geostatistics*. Springer, 2003. 18

[8] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, 2010. 75

[9] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 1987. 52

[10] D. Gallus, A. Abecker, and D. Richter. Classification of Landslide Susceptibility in the Development of Early Warning Systems. In *Proceedings of the 13th International Symposium on Spatial Data Handling*, 2008. 66

[11] D. Gallus and W. Kazakos. Einsatz von statistischen Methoden zur automatischen Erstellung von Gefährdungskarten am Beispiel gravitativer Massenbewegungen in Frühwarnsystemen. In *Tagungsband AGIT*, 2008. 66

[12] D. Gallus and M. Ruff. Classification of landslide susceptibility in the development of early warning systems. In *Map World Forum*, 2009. 66

[13] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996. 34, 49

[14] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Chapman and Hall, 1989. 56

[15] D. A. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer, 2008. 90

[16] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970. 50

[17] C. R. Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 1975. 21

[18] M. Kanevski, V. Timonin, and A. Pozdnoukhov. *Machine Learning for Spatial Environmental Data: Theory, Applications, and Software*. EFPL Press, 2009. 14

[19] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. 1933. 18

[20] D. G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical, and Mining Society of South Africa*, 52, 1951. 9

[21] N. Lawrence, M. Seeger, and R. Herbrich. Fast Sparse Gaussian Process Methods: The Informative Vector Machine. In *Advances in Neural Information Processing Systems*, volume 15, 2003. 60, 62

[22] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001. 34, 49

[23] G. Matasci. Master's thesis, Universite de Lausanne, 2009. 72

[24] G. Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 1963. 9, 20

[25] P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall, 1989. 34

[26] C. McCulloch and S. R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, 2008. 9, 10, 14, 34, 37

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. 21(6), 1953. 50

[28] T. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001. 46

[29] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2006. 31

[30] J. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In *Advances in Kernel Methods*. MIT Press. 68

[31] J. Platt. Probabilities for SV Machines. In *Advances in Large Margin Classifiers*. MIT Press. 67

[32] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C*. Cambridge University Press, 1999. 90

[33] J. Quinonero-Candela and C. E. Rasmussen. A Unifying View of Sparse Approximate Gaussian Process Regression. *Journal of Machine Learning Research*, 2005. 59

[34] J. Quinonero-Candela, C. E. Rasmussen, and C. K. I. Williams. Approximation Methods for Gaussian Process Regression. Technical report, 2007. 56, 59, 60, 61

[35] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 9, 10, 14

[36] M. Ruff. pers. communication. 67, 73

[37] M. Ruff. *GIS-gestützte Risikoanalyse für Rutschungen und Felsstürze in den Ostalpen*. PhD thesis, University of Karlsruhe, 2005. 66

[38] B. Schoelkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001. 9, 10, 14, 15

[39] S. R. Searle. *Variance components*. Wiley, 2006. 19

[40] M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations.* PhD thesis, University of Edinburgh, 2003. 62

[41] M. W. Seeger, N. Lawrence, and R. Herbrich. Efficient Nonparametric Bayesian Modelling with Sparse Gaussian Process Approximations. Technical report, MPI Biological Cybernetics, Tuebingen, 2006. 62

[42] A. J. Smola and B. Schoelkopf. Sparse Greedy Matrix Approximation for Machine Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000. 58, 59, 63

[43] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, 2006. 60

[44] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 1986. 39

[45] V. Tresp. A Bayesian Committee Machine. In *Neural Computation*, volume 12, 2000. 60, 62, 63

[46] R. von Mises. *Mathematical Theory of Probability and Statistics.* Academic Press, 1964. 88

[47] C. K. I. Williams and D. Barber. Bayesian Classification with Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1998. 43

[48] C. K. I. Williams and M. Seeger. Using the Nystrom method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, 2001. 58

Errata

p. 20: Conversely, for a finite collection $\{Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)\}$, a probability distribution, referred to as the distribution of the process, can be obtained from the stochastic process:

$p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)))$
$= \int \ldots \int p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N), Y(\mathbf{x}_{N+1}), \ldots, Y(\mathbf{x}_{N+n})))$
$dY(\mathbf{x}_{N+1}) \ldots dY(\mathbf{x}_{N+n})$,

with $n \in \mathbb{N}$, and $(\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+n})^T \in \mathbf{X}^n$.

p. 20: From above expression, $p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)))$ can be substituted for $Y(\cdot)$ if the finite collection $\{Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)\}$ is considered.

p. 20: …Hence , if interested in the distribution of finite $\mathbf{y}$, it is possible to work with the random vector $(Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N))^T$ (and the corresponding probability distribution $p(\mathbf{y}) = p((Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_N)))$, not taking $(Y_{N+1}, \ldots, Y_{N+n})$ into account.

p. 47: …by minimizing the Kullback-Leibler divergence $KL(p(y_i) \| q(y_i))$
$= KL(\int q_{-i}(y_i) p(t_i = 1|y_i) dy_i \| q(y_i))$
$= - \int q_{-i}(y_i) p(t_i = 1|y_i) dy_i \log \frac{q(y_i)}{\int q_{-i}(y_i) p(t_i=1|y_i) dy_i}$:

p. 58 $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{q=1}^{Q} \frac{\lambda_q^{(Q)}}{Q} \frac{Q}{(\lambda_q^{(Q)})^2} \mathbf{c}^T(\mathbf{x}_i) \mathbf{u}_q^{(Q)} (\mathbf{u}_q^{(Q)})^T \mathbf{c}(\mathbf{x}_j)$
$= (c(\mathbf{x}_1, \mathbf{x}_i), \ldots, c(\mathbf{x}_Q, \mathbf{x}_i))^T \mathbf{C}_{QQ}^{-1} (c(\mathbf{x}_1, \mathbf{x}_j), \ldots, c(\mathbf{x}_Q, \mathbf{x}_j))$