Predicting Bankruptcy Risk: A Gaussian Process Classification Model

Mohammed Nazib Seidu
Abstract
This thesis develops a Gaussian processes model for bankruptcy risk classification and prediction in a Bayesian framework. Gaussian processes and linear logistic models are discriminative methods used for classification and prediction purposes. The Gaussian processes model is a much more flexible model than the linear logistic model with smoothness encoded in the kernel with the potential to improve the modeling of the highly nonlinear relationships between accounting ratios and bankruptcy risk.

We compare the linear logistic regression with the Gaussian process classification model in the context of bankruptcy prediction. The posterior distributions of the GPs are non-Gaussian, and we investigate the effectiveness of the Laplace approximation and the expectation propagation approximation across several different kernels for the Gaussian process. The approximate methods are compared to the gold standard of Markov Chain Monte Carlo (MCMC) sampling from the posterior.

The dataset is an unbalanced panel consisting of 21846 yearly observations for about 2000 corporate firms in Sweden recorded between 1991 – 2008. We used 5000 observations to train the models and the rest for evaluating the predictions. We find that the choice of covariance kernel affects the GP model’s performance and we find support for the squared exponential covariance function (SEXP) as an optimal kernel.

The empirical evidence suggests that a multivariate Gaussian processes classifier with squared exponential kernel can effectively improve bankruptcy risk prediction with high accuracy (90.19 percent) compared to the linear logistic model (83.25 percent).
Acknowledgement

I would like to use this opportunity to thank the following personalities and organizations that helped throughout this thesis work.

First of all, my sincere appreciation goes to Tor Jacobson and Matias Quiroz of Sveriges Riksbank (central bank of Sweden) for providing the firm bankruptcy data. My deepest gratitude goes to my supervisor Professor Mattias Villani for spending much of your time, patience and guidance throughout the study. Your suggestions persuaded me to improve the quality of the work, it meant a lot to me, thank you sir.

I would also like to thank my family, friends and love ones for your encouragement and support, not forgetting the staff and students at the Division of Statistics and Machine Learning (STIMA) at Linköping University. It was a great experience been a student in the department, thank you all.
List of Figures

Figure 1 Graphical model of Gaussian process classifier…………………………... 11
Figure 2 Univariate Logistic and Gaussian processes models…………………………. 23
Figure 3 Laplace Approximation contour plot with SEXP…………………………….. 26
Figure 4 Laplace Approximation Joint predictive distribution with SEXP………………. 27
Figure 5 Monte Carlo integration joint predictive distribution with SEXP………………… 28
Figure 6 Latent sample draws with SEXP………………………………………………. 29
Figure 7 Automatic Relevance Determination (ARD) with SEXP……………………….. 30
Figure 8 LA contour plots with SEXP, EXP, Matern (3/2) & Matern (5/2)……… 43
Figure 9 EP contour plots with SEXP, EXP, Matern (3/2) & Matern (5/2)……… 44
Figure 10 MCMC contour plots with EXP, Matern (3/2) & Matern (5/2)……… 45
Figure 11 LA joint predictive dist. with EXP, Matern (3/2) & Matern (5/2)….. 46
Figure 12 EP joint predictive dist. SEXP, EXP, Matern (3/2) & Matern (5/2)… 47
Figure 13 MCMC joint predictive dist. with EXP, Matern (3/2) & Matern (5/2) 48
Figure 14 Latent sample draws with EXP, Matern (3/2) & Matern (5/2)……… 50
List of Tables

Table 1 Contingency table for classifying Bankruptcy risk………………………… 20
Table 2 Descriptive statistics for bankrupt and non-bankrupt firms ………………… 21
Table 3 Measuring the performance of the two classifiers using test data…………… 32
Table 4 Summarized training results from both classifiers using different thresholds….. 39
Table 5 Summarized test results from both classifiers using different thresholds……… 40
Table 6 Train data prediction results from both classifiers…………………………… 41
Table 7 Average system used in executing the covariance function………………….. 42
Table 8 Coefficient and associated p-values for the logistic regression model……… 42
1 Introduction

1.1 Background
Bankruptcy risk is an essential economic indicator in business life cycles. Historic documentations have elucidated the consequences of this event on both domestic and global trades. Due to the tremendous negative effects posed by this event, economists, practitioners and researchers have devoted much effort to study and understand the determinants and major causes of bankruptcy in small, medium and large sized corporate firms. Various countries have different definition for bankruptcy and specific authorities that declare firms as bankrupt. In most countries, the legislature enacts authorities to monitor and protect firms from failure by ensuring safe and efficient payment system. In a cashless economy, failure of one major financial institution can lead to a banking crisis that affects transactions within the payment system. When this happen, a network of firms are likely to fail because the bond in payment agreement between these firms break up causing financial instability, which may lead to a sequence of failure if it is not manage properly.

The recent financial crisis (2007-2009) indicates the extent to which a few bankrupt firms can lead to global financial distress, obviously due to the links in payment obligations and the stock exchange. As statisticians, we are interested in constructing predictive models from which we learn from accounting ratios and other financial indicators that affects firms’ failure directly or indirectly. Moreover, implementing these predictive models serves as basis for forecasting and in-depth analysis of accounting ratios that relates to bankruptcy.

Precise bankruptcy forecast models are of great importance to economists, practitioners and researchers. From a historical context many economists have proposed parsimonious economic models for forecasting bankruptcy and understanding its driving forces (Shumway 2001). The work of Altman (1968) shows how bankruptcy can be analyzed with static model. Shumway (2001), in an influential paper show how bankruptcy is
modeled using both accounting and market driven variables in a hazard model, and this model is a reference model in most econometrics studies. A hazard model allows accounting for time varying covariates and this makes it possible to use business data on different firm levels during healthy periods before default (panel datasets).

Recently, bankruptcy risk was noted to be a highly nonlinear function of commonly used explanatory variables such as firm accounting ratios (Giordani et al. 2011). They show that a nonlinear logistic model with additive splines give a much more accurate description of the data and improved bankruptcy predictions. This spline model can be used to further understand the nonlinear patterns exhibited by these accounting ratios in bankruptcy modeling. Following Jacobson et al. (2011), Giordani et al. (2011) augmented the logistic spline model with macroeconomic factors to improve the forecasting performance of the model.

In this study, we wish to explore the potential of Gaussian processes (GP) classification as a prolific model that can account for this nonlinear patterns as an alternative to splines. GPs are one of the most consistence and reliable machine learning models based on probabilistic modeling directly on distributions over functions (Vanhatalo et al. 2013). GPs have in recent years gained popularity in areas such as spatial analysis, machine learning and other scientific disciplines. A drawback of GPs is that they are computationally intensive, especially when the response variable is categorical. Nevertheless, the effectiveness of GPs in both regression and classification has motivated us to use it for bankruptcy prediction. Moreover, the results from the spline model suggest that a multivariate Gaussians would accurately describe the data and improve bankruptcy predictions. It is our intention that this work will serve as roadmap for future econometric studies regarding firm failure because the growing nature of data calls for more sophisticated flexible models. With GPs, the model complexity grows with more data and this non-parametric quality is becoming increasingly important in data-rich environments.
Our data set on Swedish corporations is an unbalanced panel data commonly used in economic studies, with information about firms in healthy periods before bankruptcy. The data set contains standard accounting ratios, macroeconomic indicators and control variables which will be used in the modeling and forecasting process.

There are several dozen covariates that one can consider using for bankruptcy risk analysis and forecasting, but the covariates used in this study are selected based on existing literature. Furthermore, the covariates reflect the highly competitive market driven economy in Sweden in relation to performance, earnings and cash flows. These financial ratios are measures of a firm’s liquidity, profitability and capital structure. GDP growth and repo rates are two macroeconomic indicators used because of their significant impact on bankruptcy predictions in recent years. Included in the models are also the control variables firm age and size, as they intuitively play a key role in understanding firms’ failure. They account for the fact that larger and older firms are often resilient to bankruptcy whilst smaller and younger firms are more vulnerable to bankruptcy.

To reach a coherent conclusion regarding a bankrupt firm, we adopt the definition used in Swedish legislation. In Sweden, the district court has the jurisdiction to declare a firm as bankrupt and this lead to the setting up of an independent liquidator to manage the assets of the firm till liquidation is completed. Creditors don’t file to court for debt settlement. After liquidation, debt settlements are made according to priority, first the cost of bankruptcy then the other creditors follows. However, if liabilities are more than assets then the government agent in charge of payment of extra debts to creditors take full responsibility for the settlement.

The remainder of the thesis is structured as follow; in the next section we present and describe the crucial parts of the bankruptcy data set. Details of the methods are elaborated in Section 3. In Section 4 we present the results for the two models (in-sample) and their predictive inferences (out-sample) with focus on the GP classification model. We discuss the results from the models in Section 5. Finally, we conclude in Section 6.
1.2 Objectives
The main focus of this thesis work is to build a Gaussian processes classifier to compute the probability of firm bankruptcy and to explore and contrast different approximation algorithms across several different covariance functions (kernels). We also investigate the factors behind bankruptcy using the Automatic Relevance Determination (ARD) criterion through MCMC for variables selection. Besides these estimations in a multiple regression setting, we also compare and contrast results from fitting univariate logistic regression and GP models to each individual variable, such an analysis gives important insights into the nature of the nonlinear relations between bankruptcy and the explanatory variables which is more easily illustrated graphically.

1.3 Definition of Terms
The basic terms used in this thesis are defined in this subsection.

Panel Data
A panel data set is observed in a multi-period setting with multiple measurements taken on each item over time. For example, suppose \( y_t^{(i)} = \{-1, +1\} \) is the dependent variable for firm \( i = 1,2, ..., n \) at time \( t \) (yearly), then this data set is considered as a panel.

Financial Ratios (FR)
FR’s are summarized information extracted from a company’s yearly report which aid in assessing the overall performance of a company. Examples of information in this statement are: net turnover, operating profit, profit after tax, earning per share etc. Financial ratios are expressed as proportion in relation to an overall size measure, often taken to be total assets.

Macroeconomic indicators (MI)
MI statistics are used by investors in analyzing and judging the performance of a company by monitoring the entire economy of a country. Investors explore macroeconomic data for current or future investment opportunities. Examples of macroeconomic indicators are: GDP growth rate, Repo rate, Unemployment data etc.

Repo rate
Repo rate is the rate at which the central bank lends money to banks (short period of time) to increase cash holding structure of these banks; this rate is also used in controlling inflation and deflation.

**Gaussian processes (GP)**
GP is a continuous stochastic process for modeling distributions over functions; it is parameterized by the mean function and a positive semi definite (PSD) covariance function $l \sim \mathcal{GP}(m(x^{(i)}), C(x^{(i)}, x^{(i)'}))$, such that $l^{(i)} = \{l^{(i)}(x_1^{(i)}), \ldots, l^{(i)}_n(x_n^{(i)})\}$ is an $n$ – dimensional vector of function values evaluated at $n$ points and $l^{(i)}$ follows a multivariate Gaussian distribution for any choice of sample points $x_1^{(i)}, \ldots, x_n^{(i)}$.

**Covariance functions (kernels)**
Covariance function is the most important object in GPs since it encodes our assumptions about the functions we are about to learn. This kernel controls the function’s smoothness and determines how function values $f(x)$ closer to each other in input space are more likely to be correlated. The squared exponential covariance function is the most commonly used covariance function.

**Latent variables**
Latent variables are not directly observed but form a crucial part in the inferential process. The function values $f(x)$ in a GP are latent variables.
2 Data

Our data set consists of 21846 yearly observations for about 2000 Swedish aktiebolag (limited liability companies) recorded over the period 1991-2008. In Sweden, all aktiebolag should by law have minimum capital of 50,000 SEK for private companies and 500,000 SEK for public companies before they are registered at Bolagsverket (Swedish Companies Registration Office - SCRO) (Bolagsverket 2014).

The data set contains both firms that file for bankruptcy after many years of operation and firms that fail already after the first few years of operation. In all, 480 firms’ entered bankruptcy in the panel, with an average failing rate of 2.20% in the entire sample period. The data is labeled such that failed firms are indicated by zero (0) whilst healthy firms are labeled by one (1). For numerical stability, we normalized the input variables to a range of [0,1] by utilizing the minimum and maximum limits for each variable. The data set is imbalanced and since the aim of classification is minimize the overall error, the minority class contributes little. There are several proposed methods to overcome this problem either at the initial stage of the data processing (for example by random oversampling without replacement) or by adjusting the decision threshold at the algorithmic level. We resort to the latter in order to reduce any potential bias.

The accounting ratios used in this study are: total liability over total assets - the leverage ratio $- \frac{TL}{TA}$, earnings before interest and tax; earnings ratio $- \frac{EBIT}{TA}$, cash and liquid assets in relation to total liability; cash ratio $- \frac{CH}{TL}$.

A firm that uses borrowed funds to buy liquid assets with the intentions of getting profit from appreciation takes greater risk and this explains the importance of leverage ratio as an indicator of bankruptcy. Concurrently, we consider companies survival based on the type of liquid assets available so that if they are short of cash, they can quickly convert these assets into money without excess lost in stock value. Companies with strong liquidity structure are less vulnerable to financial distress, also investors’ shows
substantial desire in firms with high profitability margins and for these reasons earnings and cash ratios are crucial measures of predicting bankruptcy.

The two control variables are firm size and age in years (measured from the day it was recognized as a corporate body by the Swedish authorities). Following Jacobson et al. (2011), we included GDP growth rate and repo rate as the two major macroeconomic indicators.
3 Methods
This section describes the methods used for the study. Fundamentally, classification methods are divided into two major groups namely discriminative methods where we directly model the class probabilities as a function of the variables through an activation function (example a sigmoid function) and the probabilistic generative model which uses Bayes’ theorem to invert a probability model for the variables into a posterior distribution for the class labels. Common to both approaches (models) in classification is the aim of making a precise decision threshold in feature space that separates the classes accurately, efficiently and capable of predicting new data set precisely. Also, the best models in classification are selected based on the accuracy level and other measurements (Example; Precision and false negative rate).

In this section, we describe two discriminative models for classifying and predicting bankruptcy risk, namely linear logistic regression and Gaussian processes (GPs) classification models. Intuitively, nonparametric models such as Gaussian process is expected to perform better in predictions than parametric models (linear logistic regression), because it has essentially unlimited parameters which grows with more data making it flexible in predictions. The kernels in GP model impose a certain degree of smoothness that protects against over-fitting and also improve model fits.

3.1 Probabilistic discriminative models
Discriminative methods are widely used for predicting classification problems through an activation link function which squashes its augment (data) into a range of [0, 1] making it a valid probability for the labels (the output). Generally, the goal of classification is to assign an input pattern $x$ to one of the $i$ discrete classes $C^{(1)}, C^{(2)}, \ldots, C^{(i)}$ and it’s achieved by dividing the classes into decision regions in input space, where each region is separated by a decision boundary.

3.2 Linear logistic regression model
Logistic regression models are used in applications where one is interested in estimating class probabilities for binary labels through transformation. Logistic regression is widely
used in classification and for prediction purposes, an example of such area that extensively applies this model is in credit risk screening analysis (Hastie et al. 2009). We illustrate our model by defining the following terms, for a given binary dataset $\mathcal{D} = \{(x^{(i)}, y_c^{(i)}) | i = 1, \ldots, n\}$, such that $\forall x^{(i)} \in \mathbb{R}^d$ where $d$ is the number of variables in the train data and $y_c^{(i)}$, were $c=\{0,1\}$ is the corresponding class membership for each variable, representing either a firm is bankrupt or non-bankrupt. The idea is to fit a linear logistic link function model, i.e. a logistic regression model which is linear in the variables, and apply this model in predicting the labels for a new dataset based only on knowledge of the variables. Given a vector of unknown parameters ($\omega$), the logistic regression probability model for a healthy firm ($y_c^{(i)}; c = 1$) is:

$$
p(y^{(i)} = 1) = \sigma(x^{(i)^T} \omega) = \left[1 + e^{-(x^{(i)^T} \omega)}\right]^{-1},
$$

(3.1)

where $\sigma(.)$ is the sigmoid activation function. The sigmoid function is characterized by an S-shape in space, it lies in the domain of $(-\infty, +\infty)$ mapping it inputs into a unit interval $[0,1]$ guaranteeing valid class probabilities.

### 3.3 Gaussian Processes (GPs) classification model

A Gaussian process is a continuous stochastic process that have gain popularity in machine learning applications due to its adaptive nature and flexibility in handling high dimensional data structures resulting in a consistence outcome in pattern recognition (Bishop 2006). A GPs can be viewed as probability distributions over functions because the mean and covariance are functions of the inputs ($x^{(i)}$). This model is widely used in both regression and classification analysis.

The logistic regression in Section 3.2 is based on a linear activation function $l^{(i)} = x^{(i)} \omega$. GPs generalize equation 3.1, with $l^{(i)}$ as an unknown smooth latent function. So, it is natural to place prior distributions over these smooth latent functions. For simplicity, given a set of test data $D^{(i)}_s$ , were $i = n + 1, \ldots, n_s$. We let $\mathcal{C}_{ss}$ be pairs of all data points
in the covariance matrix evaluated at \((x_*^{(i)}, x_*^{(i)\prime})\) and \(C_*\) represent the corresponding pairs of points for both train and test data \((x^{(i)}, x^{(i)\prime})\). We define a GP prior over \(l\) as

\[ l \sim \mathcal{GP}(m(x), C(x, x')) \]

where \(m(x)\) is the mean function and \(C(x, x')\) is the positive semi-definite (PSD) covariance matrix. These two quantities determine the smoothness and variability of the function. The prior latent function \(l^{(i)}\) at specific locations is generated from

\[ l^{(i)} \sim \mathcal{N}\left(m(x^{(i)}), C(x^{(i)}, x^{(i)\prime})\right) \tag{3.2} \]

The mean function \(m(x^{(i)})\) is often assumed to be zero; it does not mean in practice that the mean function is expected at zero but merely that we lack prior knowledge of the functions overall levels (Neal 1998).

\[ C \left(x^{(i)}, x_*^{(i)} | \phi \right) = \begin{bmatrix} C & C_* \\ C_*^T & C_{**} \end{bmatrix}, \text{ is the covariance matrix and it’s constructed using different covariance functions because various physical systems have different functional forms with different degrees of smoothness. } \phi \text{ is a set of hyperparameters that lives inside the covariance function and we will elaborate more on these hyperparameters in Section 3.4. } \]

With this in mind, we specify the joint prior over training \((l^{(i)})\) and test \((l_*^{(i)})\) latent values below;

\[ p(l_*^{(i)}, l^{(i)}) = \mathcal{N}\left(\begin{bmatrix} \mu^{(i)} \\ \mu_*^{(i)} \end{bmatrix} | 0, \begin{bmatrix} C & C_* \\ C_*^T & C_{**} \end{bmatrix}\right) \tag{3.3} \]

Where \(C = c(x^{(i)}, x^{(i)\prime})\) is \(n \times n\), \(C_* = c(x^{(i)}, x_*^{(i)\prime})\) is \(n \times n_*\) and \(C_{**} = c(x_*^{(i)}, x_*^{(i)\prime})\) is \(n_* \times n_*\).

For each block of this covariance matrix we decode all the entries \(C_{ij}\) (see equation 3.4) using a chosen kernel function. Furthermore, since Gaussian processes capture our belief
that most functions have different level of smoothness, we have to diverse means of capturing this belief by using the similarity measures in Section 3.4. These covariance functions are constructed such that observations close to each other are expected to have similar function values \( f(x) \) and for this reason highly correlated and those observations far away are less correlated.

\[
C(x^{(i)}, x^{(i)'}) = C = \begin{pmatrix}
C_{11} & \cdots & C_{1n} \\
\vdots & \ddots & \vdots \\
C_{n1} & \cdots & C_{nn}
\end{pmatrix}, C_{11} = C(x^{(i)}, x^{(i)'})_{11}, \tag{3.4}
\]

We model the probability of observing a healthy firm \( y_c^{(i)} = +1 \) as

\[
p(y^{(i)}|l^{(i)}) = p(y^{(i)} l^{(i)}) = \left[ 1 + e^{-l^{(i)}} \right]^{-1} = \sigma(y^{(i)} l^{(i)}), \tag{3.5}
\]

And this is how we are encoding our prior knowledge given data in Gaussian process modeling, Fig. 3.2 below shows a typical Gaussian processes classifier scheme.

![Figure 1: Graphical model of Gaussian process classifier. The target variables are independent of the other observations conditional on the latent variables. The latents are jointly modeled.](image)

3.4 Covariance Functions (kernels)

The covariance function (CF) controls the degree of smoothness of the function in the sense that nearby input observations is more likely to have similar function values. This gives Gaussian processes an extra edge over other learning models because of the natural way to impose prior smoothness via the kernel. Moreover, it encompasses these
smoothness assumptions about the data we wish to learn through a small set of interpretable hyperparameters which can be learned from data. There are several different covariance functions ranging from stationary to non-stationary, covariance functions that are invariant to translation in input space are considered stationary. Moreover if the inputs only enter the kernel in the form $||x^{(i)} - x^{(i')||}$, then the covariance function is isotropic.

One particular challenge we face in computing the covariance function is when the data set is large ($n > 10^3$ data points), we have to compute the Cholesky decomposition of large matrices, which is time consuming. GP has system run time of $O(N^3)$ for large dataset which shows that GPs takes quite some time to execute large datasets. In Rasmussen and Williams (2006) new computational methods have recently been proposed to partially solve this hurdle, but it is still one of the reasons why Gaussian processes are not commonly used in applied work. Below are some examples of commonly used covariance functions for GP classification.

### 3.4.1 Square Exponential Covariance Function (SECF)

The most commonly used covariance function is the SECF, which is of the form:

$$C(a) = \rho_{sv} \exp \left( - \sum_k \frac{(x_k^{(i)} - x_k^{(i')})^2}{2\theta_i^2} \right) + J, \quad (3.6)$$

Where, $\rho_{sv}$ is the signal variance, $\theta_i$ is a set of length scales for each variable that govern how far we move in input space before the correlation decreases, this attribute automatically determines relevant (ARD) inputs given data because the larger the length scale the more the kernel become independent of that input. $J$ are fixed “jitters” added to the kernel to prevent computational problems by contributing additively to every eigenvalues of the covariance matrix, it also prevent the covariance matrix from collapsing into singular form. $J = 1 * e^{-9}$ is used here, because large jitters draw the logit model close to that of probit model (Neal 1998). $a = (x^{(i)}, x^{(i')})$ consist of two
pairs of elements for any $x^{(i)}$ inputs. Intuitively, SECF is one of the smoothest CF that specifies the covariance between two random features.

### 3.4.2 Matérn Family Covariance Function (MFCF)

MFCF is a useful covariance kernel initially developed for spatial analysis (co-kriging) and metric space measurement. This covariance kernel has an extra positive smoothing parameter ($v$) and is of the form:

$$C(a) = \sigma^2 \frac{2^{1-v}}{\Gamma(v)} \left(\sqrt{2v\frac{a}{\theta_i}}\right)^v K_v\left(\sqrt{2v\frac{a}{\theta_i}}\right),$$  \hspace{1cm} (3.7)

The extra parameter $v$ is positive definite with $K_v$ as a modified Bessel function and $\Gamma(.)$ is the gamma function (Rasmussen and Williams 2006). A special case occurs when $v = 1/2$ (ECF) and $v \to \infty$ (SECF). The two most commonly used MFCFs in machine learning applications are:

$$C(a) = \sigma^2 \left(1 + \frac{\sqrt{3a}}{\theta_i}\right) \exp\left(-\frac{\sqrt{3a}}{\theta_i}\right) \quad v = 3/2,$$

$$C(a) = \sigma^2 \left(1 + \frac{\sqrt{5a}}{\theta_i} + \frac{5a^2}{3\theta_i^2}\right) \exp\left(-\frac{\sqrt{5a}}{\theta_i}\right) \quad v = 5/2,$$  \hspace{1cm} (3.8)

### 3.4.3 Exponential Covariance Function (ECF)

ECF is a special case of MFCF when $v = 1/2$; this stationary CF has a very rough surface in input space (Rasmussen and Williams 2006) making it unpopular in some areas of research.

$$C(a) = \rho_{sv} \exp\left(-\sum_k^d \left(\frac{x_{(i)} - x_{(j)}^{(k)}}{\theta_i}\right)^2\right) \quad J,$$  \hspace{1cm} (3.9)

### 3.4.4 GP Covariance Prior Structures

The covariance function in GPs is encoded with unknown hyperparameters and we place prior knowledge (distributions) over these hyperparameters in a fully Bayesian framework. The structure also includes fixed jitters added to the diagonal to prevent numerical problems and play a role similar to that of noise added to GP regression.
models as reported in Neal (1998). These priors represent our belief before we observed data and are often vague over the relevant domain. Also included in the prior structure are the length scales which are non-negative real numbers.

The length scale hyperparameters, one for each input measure how far we move in input space for the function values to become uncorrelated. The inverse of this length scale automatically determine how relevant an input is, with large length scale the covariance function becomes independents of that input, essentially removing it from the inference (Rasmussen and Williams, 2006). Appropriate prior densities for the positive GP hyperparameters are distributions from the exponential family; the most widely used is the inverse gamma distribution which is a continuous probability distribution on a positive real line. Here, we use the parametrization of the inverse gamma distribution often referred to as the scale-inverse-chi-square. This distribution is parameterized with two quantities, the degrees of freedom ($v$) and scaling parameter ($\tau^2 = \frac{1}{\sigma^2}$). The prior distributions for scaling $\rho_{sv}$ and length scale $\theta_i$ parameters are:

$$\rho_{sv} \sim \text{Scale} - \text{inv} - \chi^2(v_{\rho_{sv}}, \tau_{\rho_{sv}}^2),$$
$$\theta_i \sim \text{Scale} - \text{inv} - \chi^2(v_{\theta_i}, \tau_{\theta_i}^2),$$

The appropriate Bayesian joint prior for these parameters is presented below with the joint posterior distribution shown in Section 3.6.4 (Gelman 2003).

$$p(\theta, \rho_{sv}) = p(\theta|\rho_{sv})p(\rho_{sv}),$$

(3.11)

**3.5 Predictive Probability Distributions**

To draw samples from the posterior predictive distribution for the unobserved $y^{(i)}_*|i = n + 1, ..., n_*$ condition on the joint distribution of our targets cases $l^{(i)}_*|i = n + 1, ..., n_*$ which is a vector of latent functions unobserved evaluate at each input of $x^{(i)}_*|i = n + 1, ..., n_*$. The posterior predictive distribution is:

$$p(y^{(i)}_*|y^{(i)}) = \int p(y^{(i)}_*|l^{(i)}_*)p(l^{(i)}_*|y^{(i)})dl^{(i)}_*,$$

(3.12)
The posterior predictive distribution over test latent \( l^{(i)} \) is:

\[
p(l^*|y^{(i)}) = \int p(l^*|l^{(i)})p(l^{(i)}|y^{(i)})dl^{(i)},
\]

(3.13)

The posterior distribution over train latent \( l^{(i)} \) is:

\[
p(l^{(i)}|y^{(i)}) = \frac{p(y^{(i)}|l^{(i)})p(l^{(i)}|x^{(i)})}{p(y^{(i)}|x^{(i)})} \propto p(l^{(i)}|0, C) \prod_{i=1}^{n} \sigma(y^{(i)}l^{(i)}),
\]

(3.14)

where \( p(y^{(i)}|x^{(i)}) \) is the normalization constant. From the above equations we observed the following distributions \( p(l^*|y^{(i)}) \), \( p(l^{(i)}|y^{(i)}) \) and \( p(y^{(i)}|l^{(i)}) \) are non-Gaussian which has to be approximated. There are many approximation methods, our focus are on those used in GPs literature: Laplace approximation, Expectation Propagation (EP) and Markov Chain Monte Carlo (MCMC) integration.

### 3.6 Inference Algorithms

Below are the approximation methods we will use in approximating the non-Gaussian likelihood.

#### 3.6.1 Laplace Approximation (LA)

To evaluate the predictive distribution (3.12), we seek a Gaussian approximation \( q(l^{(i)}|y^{(i)}) \) to the posterior \( p(l^{(i)}|y^{(i)}) \) over the latent functions through Bayes’ theorem given by:

\[
p(l^{(i)}|y^{(i)}) = \frac{p(y^{(i)}|l^{(i)})p(l^{(i)}|x^{(i)})}{p(y^{(i)}|x^{(i)})} \propto N(l^{(i)}|0, C) \prod_{i=1}^{n} \sigma(y^{(i)}l^{(i)}),
\]

(3.15)

where \( N(l^{(i)}|0, C) \) denotes the multivariate Gaussian density prior with covariance matrix \( C \). Since we are maximizing with respect to the latent function we consider only
the un-normalized posterior. We simplify (3.15) by computing the first and second order Taylor expansion for the log likelihood and the Gaussian prior density function

\[ F(l^{(i)}) = \log p(y^{(i)}|l^{(i)}) - \frac{1}{2} l^{(i)^T} C^{-1} l^{(i)} + \text{const} \]

Hence, the first and second order derivatives are:

\[ F'(l^{(i)}) = \frac{d}{dl^{(i)}} \log p(y^{(i)}|l^{(i)}) - C^{-1} l^{(i)} \]

\[ F''(l^{(i)}) = \frac{d^2}{d(l^{(i)^2}} \log p(y^{(i)}|l^{(i)}) - C^{-1} = (C^{-1} + R)^{-1}, \quad (3.16) \]

R is the precision measure of the likelihood. At maximum \( F'(l^{(i)}) = 0 \), the mean and mode are at the same point (maximum a posterior: MAP). The expected mean function is denoted by \( \overline{l^{(i)}} \) with corresponding covariance function given by the Hessian of the negative log posterior and this can further be simplify below;

\[ \overline{l^{(i)}} = \arg \max_{l^{(i)}} p(l^{(i)}|y^{(i)}) \]

\[ R = -\frac{d^2}{d(l^{(i)^2}} \log p(y^{(i)}|l^{(i)})|_{\overline{l^{(i)}}}, \quad (3.17) \]

Also, the approximated marginal log likelihood used for learning the ARD hyperparameters is:

\[ p(y^{(i)}|\theta, \rho_{sv}) = -\frac{1}{2} \overline{l^{(i)}}^T C^{-1} \overline{l^{(i)}} + \log p(y^{(i)}|\overline{l^{(i)}}) - \frac{1}{2} \log |I + CR| \]

\[ = -\frac{1}{2} \overline{l^{(i)}}^T C^{-1} \overline{l^{(i)}} - \frac{1}{2} |I + CR|, \quad (3.18) \]

Where the first term in (3.18), is the data fit term and the second term is the complexity penalty term. By utilizing Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization
algorithm, which is an extension of Newton’s optimization method, we compute the mode and Hessian numerically. We approximate the Laplace posterior with a Gaussian distribution with mean function \( \overline{l(i)} \) and covariance matrix given by negative inverse of the Hessian in the equation below.

\[
q(l(i)|y(i)) = \mathcal{N}(\overline{l(i)}, (C^{-1} + R)^{-1}), \quad (3.19)
\]

The predictive distributions for \( q(l^*_i|y(i)) \) and \( q(y^*_i|y(i)) \) are then given by

\[
q(l^*_i|y(i)) = \mathcal{N}(C^T C^{-1} l(i), C_{i*} - C^T C^{-1} R^{-1} C_{i*}), \quad (3.20)
\]

\[
q(y^*_i|y(i)) = \int \sigma(y^*_i|l^*_i) q(l^*_i|y(i)) dl^*_i, \quad (3.21)
\]

### 3.6.2 Expectation Propagation (EP)

EP utilizes marginal moment iteratively in approximating the Gaussian posterior. To deal with the analytically intractable posterior, we approximate the likelihood using local approximation procedures through sites function \( t_i(l_i) \) (Minka 2001). Since the product of two Gaussians gives another Gaussian (un-normalized) then,

\[
q_{EP}(l(i)|y(i)) = Z_{EP}^{-1} \mathcal{N}(l(i)|0, C) \prod_{i=1}^{n} \tilde{z}_i \mathcal{N}(l(i)|\tilde{\mu}_i, \tilde{\Sigma}_i), \quad (3.22)
\]

\( \tilde{z}_i, \tilde{\mu}_i \) and \( \tilde{\Sigma}_i \) are site parameters. We minimize the Kullback-Leibler (KL) divergence between the true posterior and its approximation and then reverse this to make it analytical tractable. By solving the KL model, the parameters of the local approximating distribution (\( t_i \)) are estimated below.

\[
KL \left( p(l(i)|y(i))||q_{EP}(l(i)|y(i)) \right)
\]

\[
KL \left( q_{EP}(l(i)|y(i))||p(l(i)|y(i)) \right) = \int \mathcal{N}(l_i|\tilde{\mu}_i, \tilde{\Sigma}_i) \log \frac{\mathcal{N}(l_i|\tilde{\mu}_i, \tilde{\Sigma}_i)}{p(l(i)|y(i))} dl(i)
\]
\[ KL \left( q_{EP}(l^{(i)}|y^{(i)})|p(l^{(i)}|y^{(i)}) \right) = KL(\tilde{\mu}_i, \tilde{\Sigma}_i), \]  

(3.23)

The estimation of the local likelihood parameters \( \tilde{\mu}_i, \tilde{\Sigma}_i \) and \( \tilde{Z}_i \) are obtained from the above equation. See (Rasmussen et al. Chapter 3 pp55) for details. The approximated posterior for the EP method is

\[ q_{EP}(l^{(i)}|y^{(i)}) = \mathcal{N} \left( \Sigma \tilde{\Sigma}^{-1} \tilde{\mu}, \left( C^{-1} + \tilde{\Sigma}^{-1} \right)^{-1} \right), \]  

(3.24)

Therefore, we make prediction for distributions \( q_{EP}(l^{*(i)}|y^{(i)}) \) and \( q_{EP}(y^{*(i)}|y^{(i)}) \) given the mean and covariance matrix in (3.23) and (3.24) below;

\[ q_{EP}(l^{*(i)}|y^{(i)}) = \mathcal{N} \left( C_{**}^T (C^{-1} + \tilde{\Sigma})^{-1} \tilde{\mu}, C_{**} - C_{**}^T (C^{-1} + \tilde{\Sigma})^{-1} C_{**} \right), \]  

(3.25)

\[ q_{EP}(y^{*(i)}|y^{(i)}) = \int \sigma(y^{*(i)}|l^{*(i)}) q_{EP}(l^{*(i)}|y^{(i)}) dl^{*(i)}, \]  

(3.26)

3.6.3 Markov Chain Monte Carlo (MCMC) Integration

To evaluate the accuracy of the two approximated algorithms discussed above, we draw samples from the posterior predictive distribution (Equation 3.12) using Markov Chain techniques. Sampling from the posterior distribution for both the latent variables and hyperparameters \( p(l^{(i)}, \theta, \rho_{sv}|y^{(i)}) \) simultaneously is often difficult especially if the likelihood is non-Gaussian (Murray 2010). The most widely used Markov chains are the Gibbs and hybrid Monte Carlo sampling methods, where we sample draws for the variables of interest by alternating them condition on all the other variables in turn. These types of Markov chains take quite some time and convergence may be very slow.

Newly developed Markov chain simulation methods such as slice sampling, have proven to be faster and converge in a long run when sampling. In slice sampling, we marginalize out the latent variable and sample from the marginal posterior of \( p(\theta, \rho_{sv}|y^{(i)}) \). This makes it easier to also sample from the conditional posterior of \( p(l^{(i)}|y^{(i)}, \theta, \rho_{sv}) \) for the
latent variables using elliptical slice sampling method. We present the basic ideas in these two samplers below.

### 3.6.4 Sampling hyperparameters with slice sampler

In situations where it is often difficult to find an efficient Metropolis-Hastings proposal distribution we rely on slice sampler which does not require explicit specification. Here, we draw samples from area under the density function using sequential vertical and horizontal steps, detail of this sampler can be found in Neal (2003). We formulate the posterior distribution for the hyperparameters as

\[
p(\theta, \rho_{sv} \mid y^{(i)}) \propto p(y^{(i)} \mid \theta, \rho_{sv}) p(\theta, \rho_{sv})
\]

(3.27)

In summary, by utilizing the idea behind slice sampler, we draw posterior samples for these hyperparameters in a straightforward manner.

### 3.6.5 Sampling latent variables with elliptical slice sampler

Another robust way of sampling from the conditional predictive distribution for the latent variables \(p(l^{(i)} \mid y^{(i)}, \theta, \rho)\) is by using elliptical slice sampling algorithm. By using the elliptical slice sampler algorithm in Murray et al. (2010), we sample latent draws from conditional posterior distribution:

\[
p(l^{(i)} \mid y^{(i)}, \theta, \rho_{sv}) = \frac{1}{Z} \mathcal{N}(l^{(i)} \mid 0, C_{\phi}) p(l^{(i)})
\]

(3.28)

where \(\mathcal{N}(l^{(i)} \mid 0, C_{\phi}) = \frac{1}{\sqrt{2\pi|\Sigma|^{1/2}}} \exp \left(-\frac{1}{2} l^{(i)^T \Sigma^{-1} l^{(i)}}\right)\).

### 3.7 Linear logistic and GP classifiers

Since the final evaluation of classifiers involves assessing the performance of the different methods used, we will compare the results from these two classifiers using both accuracy, precision and false negative rate (FNR) in the contingency table shown below.
<table>
<thead>
<tr>
<th></th>
<th>Non-Bankrupt Firms</th>
<th>Bankrupt Firms</th>
<th>Marginal total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Non-Bankrupt</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
<td>Predicted Yes</td>
</tr>
<tr>
<td>Predicted Bankrupt</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
<td>Predicted No</td>
</tr>
<tr>
<td># of Non Bankrupt</td>
<td># of Bankrupt</td>
<td></td>
<td>Total # Obs.</td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{TP+TN}{Total} \)  

Precision = \( \frac{TP}{TP+FP} \)

False Negative Rate (FNR) = \( \frac{FN}{TP+FN} \)

These measurements clarify each model potential during prediction.
4 Results
In this section we divide the analysis into two main blocks namely the univariate and multivariate outputs. First, we present the factual results obtained from both univariate linear logistic (LL) and GP models in fitting the variables that relates to bankruptcy risk analysis. Then we compare our results to other similar research such as those reported in Giordani et al. (2011) spline model and Shumway (2001) gold standard model in Section 5. The second part is devoted to the results from multivariate GPs classification model with all the variables. Special emphasis is placed on comparing the results from the approximated algorithms across the several different covariance functions used in exploring the variability and smoothness of the model. Finally, we assess the models’ predictive performance using accuracy, precision and false negative rate (FNR) in Section 4.7.

In general, the descriptive statistics in Table 2, shows a clear distinction between bankrupt and non-bankrupt firms. In Table 2(a), bankrupt firms recorded negative EBIT/TA (-0.17), high TL/TA (1.02) and lower cash holding position (0.3) compared to non-bankrupt firms in Table 2(b) that shows moderate EBIT/TA ( 0.05) level, low TL/TA (0.68) and high cash holding position (CH/TL : 0.8). We present the estimated coefficients and p-values from the fitted linear logistic (LL) model in Table 8.

Table 2: Descriptive statistics for bankrupt and non-bankrupt firms from 1991-2008, EBIT/TA, TL/TA and CH/TL are the accounting ratios, whilst firm size measured in real total sales (1000 SEK) and age of firm (first registered at Bolagsverket).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Median</th>
<th>Sd</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBIT/TA</td>
<td>-0.16610</td>
<td>-0.02526</td>
<td>0.4444</td>
<td>-1.29223</td>
<td>0.76112</td>
</tr>
<tr>
<td>TL/TA</td>
<td>1.01587</td>
<td>0.92252</td>
<td>0.547410</td>
<td>0.03388</td>
<td>2.69196</td>
</tr>
<tr>
<td>CH/TL</td>
<td>0.270548</td>
<td>0.041226</td>
<td>1.04041</td>
<td>0.000000</td>
<td>12.727273</td>
</tr>
<tr>
<td>Size</td>
<td>14,024</td>
<td>14,084</td>
<td>1,5945</td>
<td>9,241</td>
<td>18,502</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
4.1 Univariate fitted models

Fig. 2 depicts the highly nonlinear patterns between bankruptcy risk and the variables of interest. The linear logistic model does not capture some of the most interesting patterns in high density regions (Fig. 1 (a)-(d)). EBIT/TA and TL/TA have similar bankruptcy risk probabilities which spans from 0 to 0.12 and firms with negative EBIT/TA faces potential risk. On the contrary, if earnings (EBIT/TA) of these firms increase, we see a monotonic decline in risk especially in high density areas and the risk stabilized afterwards (Fig. 1 (a)).

TL/TA is directly opposite to EBIT/TA in most cases, firms with lower TL/TA are relative stable at equilibrium, but as the leverage (TL/TA) of firms increase, we see a monotonic increasing relationship in regions we are more certain about the fit (areas with enough evidence). The bankruptcy risk is relatively low for firms with CH/TL level between 0.2 and 0.35 (Fig. 1(c)), we observe a plausible failure for companies with extremely low cash (CH/TL) holding position in a long run.

The patterns in Fig. 1(d)-(e) are in accordance with intuition, we expect large and old firms to be resilient to distress, but of great interest is the pattern exhibited in Fig. 1(e). Thus, we observe a gradual increase in risk for firms in operations for more 14 months

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Sd</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBIT/TA</td>
<td>0.04678</td>
<td>0.06310</td>
<td>0.2924</td>
<td>-1.29223</td>
<td>0.76112</td>
</tr>
<tr>
<td>TL/TA</td>
<td>0.68090</td>
<td>0.69117</td>
<td>0.3577</td>
<td>0.03388</td>
<td>2.69196</td>
</tr>
<tr>
<td>CH/TL</td>
<td>0.80120</td>
<td>0.23535</td>
<td>1.6686</td>
<td>0.00000</td>
<td>12.72727</td>
</tr>
<tr>
<td>Size</td>
<td>14,220</td>
<td>14,210</td>
<td>1,6496</td>
<td>9,210</td>
<td>21,650</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

(b) Non-Bankrupt Firms (Total Train Obs: 21366)
but as operations move into second quarter of second year we notice a fairly decreasing trend thereafter.

(a)

(b)

(c)
Figure 2: Univariate logistics and Gaussian processes models showing how each covariate bankruptcy risk probabilities changed over time. Each variable is grouped into equal sized quantiles making it easier to compute the bankruptcy risk probability for each sizable quantile. Quantiles are rated from low to high and firms with lower values have low quantile probability.
4.2 Multivariate GP model bankruptcy risk outputs
This part consist of the in-sample results from the different approximation algorithms across different covariance functions, so that the evaluations would be based on how the results depend on the specific combination of kernels and the approximated algorithms. Finally, we explore the prediction (out-of-sample) results from the two discriminative classifiers using an independent test data.

Using the GP model involves setting the prior hyperparameters to an acceptable level and this is achieved after reviewing many GP literatures on selecting the values that would improve the model predictive capabilities. The two parameters in our prior distribution; precision and degrees of freedom are set to 0.5 and 4 respectively. With these settings the prior density is not very informative and we expect it to combine well with our data. The hyperparameters inside the kernels, the signal variance and the lengthscales are set to 10 and 1 respectively. These values prevent the kernel from overfitting and also ensure that we don’t supply our prior with too much information such that we end up sampling from the prior distribution. A reasonable lengthscale allows the latent function to vary quickly making it possible for the classifier to use the training data in thresholding the decision surface. A shorter lengthscale leads to a more complex decision surface because it overfit the data (wiggle), while a too long lengthscales flatten the nonlinear features of the latent function due to the high level smoothness effects.

In-Sample Results
Here, we will compare the contour plots, joint predictive distributions, latent sample draws and relevant variables selection (ARD) from the three approximation algorithms. Based on our judgment from the results in fitting the GP classification model, we have chosen to present the results for the squared exponential covariance function (a standard kernel) in this section whilst the results from the other kernels can be found in Appendix A.
4.3 Laplace Approximation (LA)

Fig. 3 illustrate a contour plot from the Laplace approximation GP classifier on a binary response (bankrupt and non-bankrupt firms) using logit activation function with squared exponential kernel. Fig. 3 displays the predictive contours in EBIT/TA-TL/TA space, most bankrupt firms are classified as healthy firms and observations above the decision surface have fairly high bankruptcy probabilities.

The exponential kernel and Matern family covariance functions patterns are closely related. The decision thresholds are in accordance with our expectation (see Fig. 8 Appendix A). Moreover, the contour plots for the three approximated algorithms are rather found to be minor.

![Laplace Predictive Contour; SEXP CF](image)

Figure 3: Laplace Approximation Contour plots with squared exponential covariance function. Healthy firms are label red circle whilst blue circles represent failed firms; the black dots are the prediction for future observations. The kernel automatically determine the best decision surface, EBIT/TA and TL/TA ratios were used in constructing the contour plot.

The joint predictive distribution display an increasing trend at the extreme edges especially in areas with high EBIT/TA and TL/TA values (Fig. 4). Non-bankrupt firms are located in regions where there is a sharp descent in risk and majority of these firms have bankruptcy frequency probability less than 0.3. The predictive distribution has a distinct pattern for failed firms in north-west areas (hilly region). In general, the risk level increases when firms $EBIT/TA$ level are high and are extreme leverage $(TL/TA)$. The
Expectation propagation (EP) has similar patterns reported in Laplace approximation (LA) with the square exponential kernel, see Appendix A Fig. 9.

The joint predictive distributions for the other kernels (Exponential and Matern family CF’s) in Appendix A shows a continuous increasing trend in bankruptcy risk and spans from 0 to 0.4. Active firms have minimal low risk at the bottom of the distributions (dark blue regions). The risk is more volatile at the edges of the distributions, low TL/TA and high EBIT/TA subject firms to some level of vulnerability with probability close to 0.08 but the risk is much more profound (0.13) for firms with low EBIT/TA and high TL/TA.

![Laplace Joint Predictive Distribution For EBIT/TA and TL/TA Using SEXP CF](image)

Figure 4: Laplace Approximation (LA) Joint predictive distribution: Contour plots with squared exponential covariance function. The vertical line represents bankruptcy risk probability frequency and the diagonals are EBIT/TA and TL/TA, the colors depict the increasing trend of risk associated jointly between these two variables.

### 4.4 Markov Chain Monte Carlo (MCMC) Integration

The MCMC predictive distribution has a relative low risk at the edges and consists of mostly non-bankrupt firms (Fig.5), prediction with Monte Carlo integration is not explosive at the edges. The distribution also capture the hilly pattern at the center of the plot, risk is a bit high (0.23) were both healthy and failed firms coincide, it then increase
in full parse afterwards. Concurrently, we observed a sharp decline and a gradual tails off pattern (dark blue region) in bankruptcy risk especially in highly dense areas (Fig.5).

The Matern family kernels show more flexible and smoother contour lines with distinct separation (decision boundary) at 0.15 (See Appendix A Fig.10). Fig.10 Appendix A, illustrate the joint predictive distributions for Exponential and Matern family kernels with Monte Carlo integration, we detect a moderate smoothness in both high and low zones of the distributions, notice the risk level at the lowest edges of the plots.

Figure 5: MCMC Joint predictive distribution: Contour plots with squared exponential covariance function. The vertical line represents bankruptcy risk probability frequency and the diagonals are EBIT/TA and TL/TA, the colors depict the increasing trend of risk associated jointly between these two variables.

4.5 Random draws for two latent variables in input locations
Fig.6 illustrates a draws for two latent variables at some randomly selected locations. The gold standard algorithm (blue histogram) is symmetric and the two approximation procedures fit the distribution quite well (first plot from left). Both LA and EP distributions are close to the posterior mass with a bell shape characteristics. The second plot is a bit skewed from the posterior mass for both LA and EP, but at least the variation seems to be approximated very well.
We observe a shift in LA distribution towards the posterior mode whilst EP is centered at the mass (mean) of the distribution with exponential kernel (Fig.14 (a) Appendix A). The first latent draws for both exponential and Matern family kernels are close to normal compare with the second draws which is skewed and lack the qualities of a bell-shape distribution (Fig.14 (a)-(c) Appendix A).

Figure 6: Random Sample Draws for latent variable at different input locations using squared exponential covariance function. The legend defines the three different algorithms used.

4.6 Automatic Relevant Determination (ARD) with MCMC

The hyperparameters were sampled using Monte Carlo integration through slice sampling, by plotting the values of the hyperparameters we assess how they change during the course of simulation and it gives us more insight into the decaying patterns of the length scales in the exponential part of the kernel. In Fig.7 (a)-(b), we observe the improvement of the hyperparameters (\(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6 \& \theta_7\)) in the squared exponential covariance over 120 iterations. The Markov chain convergence is slow and the number of iterations is not sufficient due to limited time. But, what we wish to see in
the simulation are the length scales with relatively low values. In general, there are fluctuations in the draws due to choice of sampler and it is a trade-off between speed and accuracy. However, on average we observe high length scale values for $\theta_1$ and $\theta_3$ in Fig.7 (a) associated with relevant inputs whilst $\theta_2$ have small values comparatively. Also, in general the weights for $\theta_4$, $\theta_6$ and $\theta_7$ are quite high compare to that of $\theta_5$ in Fig.7 (b).

Figure 7: Relevant hyperparameters sampling using squared exponential covariance function, LS1, LS2 and LS3 are the financial ratios in order EBIT/TA, TL/TA and CH TL. The other ratios are LS4, LS5, LS6 and LS7 representing; firm size, age, GDP growth rate and repo rate. Number iterations;120 with 20 burn-in

4.7 Out-of-Samples (Prediction) Results
Table 4-5 in Appendix A shows the summarized training and test results for both classifiers using different decision thresholds that minimize the penalty in case of an incorrect classification. In practical situations, the optimal decision boundary can only be
determined through decision theory where the expected loss function plays an essential role. The optimal threshold is normally determine by the central bank by looking at both the positive and negative consequences in giving loans to corporate firms classified as false positive or false negative cases. We present the computed measures used in assessing the predictive performance of the two discriminative classifiers in Table 3. In Table 6 in Appendix A, the GP model performed quite well compared to the linear logistic model though both models are sensitivity to the imbalance nature of the training data.

The prediction results for the classifiers are highlighted in Table 3 using the same decision thresholds with the corresponding confusion matrices in Table 4 - 5. The results reported in Table 3 clearly favor the GP model. The precision level for these two model are similar, however the accuracy using the two decision thresholds is high for the GP model compared to the linear logistic model (see Table 3). Furthermore, the GP model have minimal false negative rate (7.91%) compare to the linear logistic model (14.92%) in all the decision thresholds used.

Table 3: Prediction performance of the two classifiers using balance test data (16846)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Logistic Classifier (%)</th>
<th>GP Classifier (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>0.991</td>
<td>0.0551</td>
</tr>
<tr>
<td>Accuracy</td>
<td>83.25</td>
<td>89.71</td>
</tr>
<tr>
<td>Precision</td>
<td>97.45</td>
<td>97.74</td>
</tr>
<tr>
<td>FNR</td>
<td>14.92</td>
<td>8.39</td>
</tr>
<tr>
<td>Threshold</td>
<td>0.990</td>
<td>0.0571</td>
</tr>
<tr>
<td>Accuracy</td>
<td>78.11</td>
<td>90.19</td>
</tr>
<tr>
<td>Precision</td>
<td>97.36</td>
<td>98.00</td>
</tr>
<tr>
<td>FNR</td>
<td>20.23</td>
<td>7.91</td>
</tr>
</tbody>
</table>
5 Discussion
The findings explicitly suggest a Gaussian process model is a consistence method in classifying and predicting bankruptcy risk with high accuracy (90.19 percent) compared to the linear logistic model (83.25 percent) using the decision thresholds. The main challenge for using GPs classification for bankruptcy prediction is the system run time (speed) for the parameter inferences. On average, the minimum time used to execute all runs for each kernel was closed to ten thousand minutes; the squared exponential covariance function was the fastest and the other kernels used almost three times more CPU time compare to the squared exponential kernel.

We explored many options to improve the predictive performance of the linear logistic and in particular the Gaussian processes classification models. Throughout the study the idea was to apply statistical techniques in determining the nonlinearity characterized by selected financial ratios and further explore the potentials of both classifiers in forecasting bankruptcy risk. We explore and compare the effectiveness of various widely used algorithms to approximate the non-Gaussian posterior distribution. Furthermore, four covariance functions with different levels of variability and smoothness were incorporated to improve the quality of the classification processes and also capture the nonlinear patterns pertaining to each variable of interest in the GP model.

In Table 8, the p-value statistic suggests earnings, leverage and firm age are the only important bankruptcy risk predictors in the LL model. The univariate relationships between the variables and bankruptcy risk in Fig.2, confirms the effectiveness and flexibility of the GP model characterized by the nonlinear latent function which significantly improve the model fit and also capture the time dependency nonlinear patterns in the data compared to the linear logistic model. The univariate GP fits in Fig.2 exhibit monotonic nonlinear relationship with bankruptcy risk. Moreover, the GP fitted model results are in line with the spline model results reported in Giordani et al. 2011.
The differences between the two approximation methods, Laplace (LA) and Expectation propagation (EP), are shown to rather minor. The type of kernel has a larger effect on the results. In Fig.3, the LA Contour plot with squared exponential covariance function shows moderate smoothness, the other kernels used in LA approximation shows smooth surfaces but misses some of the most interesting patterns. Contrary to the squared exponential (SEXP) kernel, the joint predictive distributions for the exponential (EXP), Matern (3/2) and Matern (5/2) kernels give bankruptcy risk probabilities increments that are monotonic in all directions compared.

One useful comparison of the approximation algorithms is to depict the posterior distributions of the latent function at given input. MCMC sampling is the gold standard algorithm because it gives the exact posterior distribution from simulation and for this reason we superimpose the other two approximations on the posterior distribution of the gold standard algorithm. In this research we sampled two draws each for the kernels, in Fig.8 the squared exponential kernel Gaussian approximations has features closed to a normal distribution (bell-shape). Both LA and EP sample draws are close to the posterior mass of the distribution especially the EP since the optimization under this algorithm is approximated by local site functions (mean & sigma) and not the mode (MAP), so the EP optimizer search for the posterior mass of the distribution. The Exponential, matern (3/2) and matern (5/2) covariance functions for the Gaussian approximations are skewed and the worse of it all are in the second draws for these kernels which lack normality features compared with squared exponential covariance function draws (see Fig.14).

The Automatic Relevance Determination (ARD) iterations in Fig.7 using kernel in equation 3.4 shows how the hyperparameters in the exponential part change over time with slice sampler. Alternative to the direct Markov chain simulation is to optimize the log-marginal likelihood using LA procedure before integrating over the entire distributions (Equation 3.16 ), although work by Minka (2001) demonstrates the effectiveness of expectation propagation ARD over LA optimization procedure. Another
challenge in this study was in relation to the type of sampler to use, Markov chains that are accurate to use take very long time to converge.

The simulations convergence are slow especially with fewer iterations but what we want to demonstrate in this study is to show the decaying patterns of these length scales during the course of simulation, so we compare on average the length scales with low values in the iteration and these are associated to irrelevant inputs. The Markov chain has a lot of fluctuations and it is a price we pay for the trade-off between model accuracy and system run time (speed). In generally, we are pleased with the ARD draws on average using the kernel in equation 3.4. We notice low length scale weights for $\theta_2$ and $\theta_5$ during the cause of simulation and for this reason they are associated with irrelevant inputs relative to $\theta_1, \theta_3, \theta_4, \theta_6$ & $\theta_7$ (see Fig. 7). The relevant variables in forecasting bankruptcy risk using Gaussian processes are: EBIT over total assets (earnings ratio), cash and liquid assets in relation to total liabilities (cash ratio), firm size, GDP growth and repo rate whilst the irrelevant variables are; total liabilities over total assets (leverage ratio) and firm age though we can further improve this simulation for convergence to be reached.

The important variables in GP model are in contrast to the linear logistic model except earnings ratio that was found as a common variable among these two models. So basically, the GP model identified more extra important bankruptcy risk predictors compared to the linear logistic model. In comparison (GP model) with Shumway (2001), the leverage ratio was reported as an irrelevant bankruptcy risk predictor and firms that are highly leveraged are still active although faces some level of insolvency distress. The out-of-sample results for the classifiers convey the result that a nonlinear decision surface model improves the classification performance compared to a fixed linear decision boundary model. Moreover, the false negative rate (FNR) for the GP model (7.91%) is approximately two less compared to the linear logistic model (20.23%).

A main contribution in this thesis is a thorough comparison of the performance of different approximation algorithms for different choices of covariance functions. The
Laplace (LA) and Expectation propagation (EP) performed remarkable well during the approximations processes and this further reflect why they are extensively used in Gaussian process modeling. Both approximates posterior distributions were close to the posterior mass of the MCMC distribution with a bell shape characteristics.
6 Conclusions
This paper develops a Gaussian processes model for bankruptcy risk classification and predictions using standard covariance functions. The GP model is a much more flexible model compared to the linear logistic model because of its nonparametric nature with guards against overfitting coming from the assumptions of smoothness encoded in the kernels to improve the model predictive performance. Our findings can be summarized into four main categories.

Firstly, the GP model allows nonlinear and nonparametric relationships between financial ratios and bankruptcy risk compared to the much more rigid linear logistic model. Secondly, an important finding is that the GP model strength lies within the covariance function used to gauge the model prediction performance and we find support for the squared exponential covariance function (SEXP) as an optimal kernel.

Next, the relevant bankruptcy risk predictors according to the MCMC inferences on the ARD parameters are: EBIT over total assets, cash and liquid assets in relation to total liabilities, firm size, GDP growth and repo rate. The other two indicators total liabilities over total assets and firm age are not significant predictors of bankruptcy which is in contrast to the linear logistic model results.

Finally, both the in-sample and out-of-sample outcomes in the classification process explicitly advocate our endorsement of the GP model in this paper. Moreover, the GP model outstanding performance with respect to accuracy (90.19 percent) makes it a top contender for the best available bankruptcy risk classification and prediction model.
Further Works

Due to the computational intensive nature of the GP model and the fact that we received the second batch of the data late, we used the first set received (5000) to train our models and the rest for testing since it is much easier to perform prediction using large data set. Another hurdle that can be look at in future work in relation to this type of classification is the imbalance nature of the data set and this has much influence in the classification process because the minority class contributes little. Both models are sensitive to this property (imbalance), but it is obvious because we expect an adaptive model to behave in such manner. Basically, there several ways to deal with this imbalance property of the data set, first at the initial data processing level or secondly at the algorithmic level were we adjust the decision threshold.

In the data processing level Ganganwar (2012) propose sampling techniques were we increase the size of the minority class (oversampling) although this method replicates the existing examples in the minority class, overfitting is more likely to occur. On the contrary, we can also decide to sample subset of the majority class (undersampling) for training although this approach is also inefficient in classification process (Ganganwar 2012). Another remedy is to use the methods proposed by Rasmussen and Williams (2006) in handling large data size.
Literature


Ghahramani, Zoubin, Class Tutorial, Topic, *Gaussian Processes (or why I don’t use SVMs)*, Department of Engineering, University of Cambridge, UK, 2006.


Appendix A: Tables and Figures

Table 4: summarized training results (5000 observations) from both classifiers with decision surfaces of 0.991, 0.05511, 0.99 and 0.0571 for linear logistic and GP classifiers outputs

(a)

<table>
<thead>
<tr>
<th>Decision Line=0.991</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>4183</td>
<td>97</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>718</td>
<td>2</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Decision Line =0.05511</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>4488</td>
<td>92</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>413</td>
<td>7</td>
</tr>
</tbody>
</table>

(c)

<table>
<thead>
<tr>
<th>Decision Line=0.990</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>3941</td>
<td>95</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>960</td>
<td>4</td>
</tr>
</tbody>
</table>

(d)

<table>
<thead>
<tr>
<th>Decision Line =0.0571</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>4512</td>
<td>92</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>389</td>
<td>7</td>
</tr>
</tbody>
</table>
Table 5: Prediction results (16846 test data) from both classifiers with decision surfaces of 0.991, 0.05511, 0.99 and 0.0571 linear logistic and GP classifiers outputs

(a)  
<table>
<thead>
<tr>
<th>Decision Line = 0.991</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>14009</td>
<td>366</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>2456</td>
<td>15</td>
</tr>
</tbody>
</table>

(b)  
<table>
<thead>
<tr>
<th>Decision Line = 0.05511</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>15083</td>
<td>349</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>1382</td>
<td>32</td>
</tr>
</tbody>
</table>

(c)  
<table>
<thead>
<tr>
<th>Decision Line = 0.990</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>13134</td>
<td>356</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>3331</td>
<td>25</td>
</tr>
</tbody>
</table>

(d)  
<table>
<thead>
<tr>
<th>Decision Line = 0.0571</th>
<th>No Default</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted No Default</td>
<td>15162</td>
<td>350</td>
</tr>
<tr>
<td>Predicted Default</td>
<td>1303</td>
<td>31</td>
</tr>
</tbody>
</table>
Table 6: Measuring the performance of the two classifiers using training data with decision surface for both LL and GP models respectively.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Logistic Classifier (%)</th>
<th>GP Classifier (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>0.991</td>
<td>0.05511</td>
</tr>
<tr>
<td>Accuracy</td>
<td>83.70</td>
<td>89.90</td>
</tr>
<tr>
<td>Precision</td>
<td>97.77</td>
<td>97.80</td>
</tr>
<tr>
<td>FNR</td>
<td>14.65</td>
<td>8.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>Logistic Classifier (%)</th>
<th>GP Classifier (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>0.990</td>
<td>0.0571</td>
</tr>
<tr>
<td>Accuracy</td>
<td>78.90</td>
<td>89.72</td>
</tr>
<tr>
<td>Precision</td>
<td>97.65</td>
<td>97.77</td>
</tr>
<tr>
<td>FNR</td>
<td>14.92</td>
<td>8.39</td>
</tr>
</tbody>
</table>

Table 7: Average system used in executing the task, we used a normal desktop top in the university statistics lab which has a processor speed of 3.30 GHz.

Statlab Machine (3.30 GHz Intel(R) Xeon(R) CPU E31245 Processor)

<table>
<thead>
<tr>
<th>Covariance Function</th>
<th>Time (Minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEXP</td>
<td>4433</td>
</tr>
<tr>
<td>EXP</td>
<td>12305</td>
</tr>
<tr>
<td>Matern(3/2)</td>
<td>12359</td>
</tr>
<tr>
<td>Matern(5/2)</td>
<td>11798</td>
</tr>
</tbody>
</table>

Table 8: Parameters estimates and associated p-values for the logistic regression model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Standardized Estimated Slope ($\beta_i$)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.7697</td>
<td>0.000</td>
</tr>
<tr>
<td>EBIT/TA</td>
<td>1.6247</td>
<td>0.016</td>
</tr>
<tr>
<td>TL/TA</td>
<td>-2.5872</td>
<td>0.000</td>
</tr>
<tr>
<td>CH/TL</td>
<td>3.0348</td>
<td>0.181</td>
</tr>
<tr>
<td>Size</td>
<td>1.0619</td>
<td>0.229</td>
</tr>
<tr>
<td>Age</td>
<td>-0.9403</td>
<td>0.000</td>
</tr>
<tr>
<td>GDP</td>
<td>-0.1262</td>
<td>0.830</td>
</tr>
<tr>
<td>repo</td>
<td>-0.1915</td>
<td>0.8045</td>
</tr>
</tbody>
</table>

(a)
Figure 8: Laplace Approximation Contour plots using different level of smoothness (kernels). From (a)-(c) are the Exponential covariance function, Matern (3/2) covariance function and Matern (5/2) covariance function respectively. Healthy firms are label red circle whilst blue circles represent failed firms; the black dots are the prediction for future observations. Each kernel created its own decision boundary based on the smoothness and data information, EBIT/TA and TL/TA were used in constructing the contour plots.
Figure 9: Expectation Propagation contour plots using different kernels, from (a)-(d) are the Squared Exponential kernel, Exponential covariance function, Matern (3/2) covariance function and Matern (5/2) covariance function respectively. Healthy firms are label red circle whilst blue circles represent failed firms; the black dots are the prediction for future observations. EBIT/TA and TL/TA were used in constructing the contour plots.

(a)

(b)
Figure 10: MCMC contour plots using different kernels, from (a)-(d) are the Squared exponential kernel, exponential covariance function, Matern (3/2) covariance function and Matern (5/2) covariance function respectively. Healthy firms are label red circle whilst blue circles represent failed firms; the black dots are the prediction for future observations. EBIT/TA and TL/TA were used in constructing the contour plots.
Figure 11: LA Joint predictive distributions using different kernels, from (a)-(c) are the Exponential covariance function, Matern (3/2) covariance function and Matern (5/2) covariance function respectively. The vertical line represents bankruptcy risk probability frequency and the diagonals are EBIT/TA and TL/TA, the colors depict the increasing trend of risk associated jointly between these two variables.
Figure 12: EP Joint predictive distributions using different kernels, from (a)-(d) are the Squared Exponential kernel, exponential covariance function, Matern (3/2) covariance function and Matern (5/2) covariance function respectively. The vertical line represents bankruptcy risk probability frequency and the diagonals are EBIT/TA and TL/TA, the colors depict the increasing trend of risk associated jointly between these two variables.
Figure 13: MCMC Joint predictive distributions using different kernels, from (a)-(c) are the Exponential covariance function, Matern (3/2) covariance function and Matern (5/2) covariance function respectively. The vertical line represents bankruptcy risk probability frequency and the diagonals are EBIT/TA and TL/TA, the colors depict the increasing trend of risk associated jointly between these two variables.
Figure 14: Random Sample Draws for latent variable at different input locations, from (a)-(c) are the Exponential covariance function, Matern (3/2) covariance function and Matern (5/2) covariance function distributions respectively. The legend defines the three different algorithms used.