Stockholms universitet

Exotic approaches for modelling Loss Given Default

Felix Martinsson

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2017:10
Matematisk statistik
September 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2017:10**
http://www.math.su.se

Exotic approaches for modelling Loss Given Default

Felix Martinsson*

September 2017

Abstract

One of the main risks for a commercial bank is the credit risk, the risk that the counterparties won't pay back their outstanding amount. From an institutional as well as a regulatory perspective, this creates the need for proper statistical methods for modelling credit risk. The ambition of this thesis was to dig into and develop statistical theory behind Loss Given Default (LGD) modelling, one of the main components of credit risk. In particular, LGD was looked upon from an IFRS 9 perspective, which is a new global, regulatory standard for handling credit risk from an accounting viewpoint. Two types of approaches for modelling LGD were investigated in particular. The first one was based on standard regressions extended to include a time varying intercept represented as a latent variable. The rationale was to improve the handling of the time serie dimension in the data. The second approach was to model LGD with the machine learning methods Support Vector Machine Regression. These approaches were applied on data from a Swedish corporate portfolio. The empirical results were inconclusive, but with some support of the usefulness of the approaches. From a theoretical view, the approaches seemed to have potential. More research as regards theoretical development as well as practical application are needed to further improve LGD modelling.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: felix.martinsson@gmail.com. Supervisor: Chun-Biu Li.

Acknowledgement

I'm grateful to my academic supervisor Chun-Biu Li, from Stockholm University, for support regarding the theoretical understanding of the models throughout this thesis, and to my corporate supervisor Hanna Wu, from Nordea, for help regarding knowledge of LGD, IFRS 9 and the modelling data. Furthermore, I'm thankful for my employment at Nordea on which I've learnt a great deal about LGD and loss data which has created a lot of synergies with this thesis. At last, I would like to thank my girlfriend Hana Bajrami for supporting me throughout my master studies.

Contents

1 Introduction

This section will describe the problem background of this thesis and continue with the purpose and statistical approaches that will be employed for the modelling of LGD.

1.1 Background

Every loan brings an uncertainty, that is the risk that a counterparty to a financial institution, be it an individual or a corporation, will not pay back the agreed amount when the loan is due, and that the internal process and external debt collection authority fails to collect the money. If a bank has a loan, or an *exposure*, which is the more formal term, of EUR 100.000, it represents an asset for the bank and a liability for the counterparty. But there is more to the contract than the face value of the loan. The lender has to pay interest, which is the cash flow booked as an asset for the bank. In addition to the interest rate cash flow, there is also the risk of full or partial depreciation of the exposure if the counterpart fails to repay the bank. This is a stochastic cash flow, referred to as *credit risk*.

Credit risk is built up of several components. The first component is the default event D, that is an indicator random variable that indicates whether an exposure is in default.[1] The expected value of this indicator is the probability that a business partner will go into default, referred to as *Probability of Default*, or PD.

The second one of the main credit risk components is the *Loss Given Default*, LGD, that is how much of the exposure that will be lost given a default. The unit is percentage, meaning that a LGD of 50% says that half of the exposure value has been lost. The third component is the *Exposure at Default*, EAD, which is the value of the exposures belonging to a customer at the default time. The unit is in real currency, meaning that an EAD of EUR 1.000 says that the total exposure value of the customer at time of default was EUR 1.000. For many exposures, EAD will be deterministic, however for exposures like credit cards, where the debt changes according to the will of the customer, the value at default is stochastic.

Combining the credit risk components yields the formula for credit loss in Equation 1.

$$Loss = D \cdot LGD \cdot EAD \tag{1}$$

[1] The definition of default in this context is that the counterparty has applied for bankruptcy, indicated that it will not pay back the loan after it is due, or that the payment has been 90 days past due date.

The expected value of the loss in Equation 1 under the conventional assumption that the LGD and EAD are uncorrelated in Equation 1, yields the standard formula for the *Expected Loss* in Equation 2, which is what the bank can expect to lose from the customer.

$$\textit{Expected Loss} = PD \cdot E(LGD) \cdot E(EAD) \tag{2}$$

International Financial Reporting Standard (IFRS) 9 is an upcoming, global accounting standard which addresses the accounting of financial instruments. IFRS 9 more or less concerns the calculation of the expected loss in 2 and how to account for this in the *balance sheet*, which is a main financial report that lists all assets and liabilities and their value. This thesis will be focused on modelling LGD in an IFRS 9 context.

1.2 Method

Four different approaches will be set up for the modelling of LGD. The first one is a standard regression approach, while the second extends the first by including time as a continuous random effect. The third and fourth are machine learning approaches, making use of the *Support Vector Machine* and *Support Vector Regression* models. All the approaches will partion LGD into a probability for a non-zero LGD, and a conditional LGD given that it is non-zero, and model these separately. These components, referred to as *loss probability*, and *expected conditional severity*, or just *severity*, will be modelled separately for all approaches, meaning all approaches will have a model for estimating the probability of loss and another one for estimating the conditional severity. The partition of LGD is shown in Equation 3.[2]

$$\begin{aligned} E(LGD) &= E(LGD \cdot 1_{LGD=0} + LGD \cdot 1_{LGD\neq 0}) = E(LGD \cdot 1_{LGD\neq 0}) \\ &= E(E(LGD \cdot 1_{LGD\neq 0}|1_{LGD\neq 0})) = E(1_{LGD\neq 0} \cdot E(LGD|1_{LGD\neq 0})) \\ &= \underbrace{E(1_{LGD\neq 0})}_{\text{Probability of loss}} \cdot \underbrace{E(LGD|1_{LGD\neq 0})}_{\text{Conditional severity}} \end{aligned} \tag{3}$$

The four approaches that will be employed are:

- **Approach 1**: *Modelling probability of loss with a logistic regression, modelling expected severity with a linear regression*
- **Approach 2**: *As model 1, but includes time as a continuous-time random effect in both the linear and logistic regression*
- **Approach 3**: *Modelling probability of loss with a Support Vector Machine extended with Platt's probability scoring, modelling expected severity with a Support Vector Regression and including time as an explanatory variable*

[2]The indicator variable will here and throughout the thesis be denoted as $1_{[\cdot]}$, which equals one if the condition in the subscript is true and otherwise zero.

- **Approach 4**: *Modelling probability of loss with a Support Vector Machine extended with Platt's probability scoring, modelling expected severity with a Support Vector Regression and excluding time as an explanatory variable*

1.3 Purpose

There are many studies regarding IFRS 9 and credit risk in general, but this will take on a more specific role with a special focus on using Approach 2-4 for modelling of LGD. This thesis will be more focused on the mathematics behind the approaches, and their success in modelling of LGD, and less on viewing LGD from a business, data and accounting perspective. It will give some background to why the LGD is of interest from the regulatory IFRS 9 perspective, and especially try to make the model handle time dependency as well as possible, since it is one of the main goals with IFRS 9, which will be elaborated on later.

The main question of the thesis is to compare Approach 1 with Approach 2-4 and answer the question whether the data support the use of advanced approaches instead of a standard one. Approach 1 has been chosen as a benchmark in order to test whether the more advanced approaches improve the results. Approach 2 has been chosen as an extension of Approach 1 to see whether a more thoroughly time dependent model improves the prediction in data in a different time period than the training data. Approach 3 and 4 have been chosen due to their successful use in several papers which will be discussed in the coming section. The reason for having both Approach 3 and 4, which are very similar since both use SVM and SVR, is to see how the time dependency is handled in Approach 3. There are time dependency in LGD, but it is not obvious how this will be handled in different approaches, especially with SVM and SVR since they are non-linear and hard to interpret. The outcome when time is included explicitly as an explanatory variable is uncertain, and therefore both Approach 3 and 4 are set up.

A secondary purpose of this thesis is to explore the self-developed model that lays behind Approach 2, which I hope can be relevant for other data as well.

2 Credit Risk

This section will present a background on credit risk and the modelling of it, and put LGD into a financial context. It will especially describe IFRS 9, an upcoming regulatory framework where LGD plays a central role.

According to the *Basel Committee on Banking Supervision*, an international committee where central bank governors of more than 30 countries are represented, credit risk is defined as the *[...] potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms.*[3]

Credit risk is the main risk commercial banks are facing which brings the need for employing proper statistical models for the estimation and monitoring of it. Credit risk is of interest for several perspectives in a bank including for

- pricing loans by requiring an interest rate which will cover the cost of bearing the risk.
- calculation of the *regulatory capital requirement*, which is the legal requirement of the capital that a bank must hold in order to be able to cover stressed loan losses in case of a credit crisis.
- adjusting the balance sheet so that the outstanding loans are evaluated to their actual value.

This thesis is concerned with the last of these perspectives, which is referred to as *impairment accounting*.

2.1 Impairment accounting

The international standards for financial accounting are the *International Financial Reporting Standards* (IFRS) which are reported by the *International Accounting Standards Board* (IASB). The IFRS standards are under constant development and from time to time an old standard is replaced by a new one. This is the case with IFRS 9, which will replace IAS39 as of 1th January 2018. IFRS 9 concerns classification and measurement of financial instruments, hedge accounting and impairment of financial assets, of which the last contains the most important changes and which this thesis will concern.

A balance sheet consists of assets on one side and equity and liability on the other. For a commercial bank, the most important assets are in general the exposures. This valuation of those assets may theoretically be done in many ways going from the simplest way of the exposure amount at present day to a more complicated way of the discounted expected value taking into account interest rates as well as the expected value of losses coming from a possible default. The expected value of the losses coming from a default should be accounted

[3] Basel, *Principles for the management of credit risk*.

for in the balance sheet. This is done by so called *provisions*, which is a post in the balance sheet that should cover credit losses. For customers in default, an individual provision is assessed which takes into account individual factors about the specific customer. For customers not in default, the provisioning is more reliant on statistical modelling.

Like Basel III and most other financial regulations in the past decade, IFRS 9 is born from the Great Recession. One of the learnings from that crisis from a transparent point of view was that the value of the exposures at the balance sheet were not reevaluated even though the value had decreased. This has caused two of the main requirements of the IFRS 9 that the models for predicting credit loss shall be *forward looking* and *point in time*.

In the IFRS 9 context, forward looking means that the models should take future predictions of change in the economy into account in the predictions of future losses. The requirement of point in time means that the provisions should match what the bank expects to lose due to credit losses in the current portfolio. This means that the provisions should be higher in the midst of a crisis than in a boom. This is in contrast with the requirements of the regulatory capital requirement, where the modelling should not take the state of the economy into account, also referred to as *over-the-cycle* modelling.

2.2 Modelling of LGD

Even though credit risk modelling has existed for several decades and has had a big surge in its popularity due to the release of the Basel II regulation in 2004, which gave incentives to credit risk modelling since it could decrease capital requirement, most of the focus has been on modelling Probability of Default rather than on LGD.[4] It has been, and still is, very common with LGD modelling that is based on simple averages for different buckets.[5] However in the last years, more and more light has been shed on LGD.

Machine learning has been used for LGD modelling at least since 2004, when L. Allen, G. DeLong and A. Saunders applied artificial neural networks for LGD modelling with positive results.[6] Furthermore, the last decades surge on machine learning indicates that its presence in credit risk modelling will only grow further.

As of now, there have been just a few studies where SVM or SVR have been utilised for the modelling of LGD. The first study seems to have been done by G. Loterman et al in 2012.[7] They compared 24 regression techniques, including SVR, for modelling LGD on data from five financial institutions and found

[4] Loterman et al., 'Benchmarking regression algorithms for loss given default modeling'
[5] Gupton et al., 'LOSSCALCTM: Model for predicting loss given default (LGD)'
[6] Allen et al., 'Issues in the credit risk modeling of retail markets'
[7] Loterman et al., 'Benchmarking regression algorithms for loss given default modeling'

that non-linear methods such as SVR, but also neural networks, had the best performances.

X. Yao, J. Crook and G. Andreeva also did an extensive benchmark study of LGD modelling and compared 13 approaches for modelling LGD in 2015. These included SVR, fractional response regression and linear regression, and several types of transformations, such as beta transformation were tried.[8] They found that SVR performed significantly better than the other approaches, but that transformations, for SVR as well as for linear regression, did not improve the fit.

E. Johnston Ross and L. Shibut from FDIC did a study of different LGD models in 2015 and found that a loss given loss approach, i.e. dividing LGD in the indicator whether LGD will be greater than zero, and the conditional value given that it is greater, was a good approach.[9] This approach also seems reasonable from a data perspective, since a large portion of the LGDs in general stay at zero.

Both G. Loterman et al and X. Yao et al used a SVR on the whole LGD dataset, without dividing LGD into a loss indicator component and a loss severity component. This thesis will therefore use a new approach since it combines the SVR with a loss given loss approach. The hope is that this combination of modelling approaches will add value regarding using SVM and SVR for modelling LGD.

[8] Yao et al., 'Support vector regression for loss given default modelling'

[9] Johnston Ross and Shibut, 'What Drives Loss Given Default? Evidence from Commercial Real Estate Loans at Failed Banks'.

3 Data

The distribution of LGD differ between countries, institutions and the products in the credit portfolio, but LGD is generally characterised by a bimodal shape with a large portion of the defaults bringing no loss at all, since the counterparty manages to settle the debt, and a minor peek at LGD equal to one, meaning that all outstanding debt had to be written off. The bulk of the rest of the distribution lays between zero and one.[10] It is possible with negative values, and values greater than 1. Values greater than one can occur for example if an additional drawing is allowed after defaults, with the purpose of getting a company to be performing again, but the whole amount plus the new drawing is fully lost anyhow. However, LGDs outside the range $[0, 1]$ tend to be uncommon, and the ones less than zero will be ceiled to zero in this thesis.

One of the most important factors that drives LGD is *collaterals*, which are assets pledged for covering the exposures.[11] In a retail setting, mortgages tends to be secured by real estate collateral, which is also common for corporates. Other type of collaterals includes cars and factories. Similar to collaterals in the sense that it is securitisation of the loan are *guarantees*. It could be that a parent guarantees the mortgage of his child, which makes the parent legally obliged to pay the interest and amortisation should the child not. It could also be a parent company that guarantees the loans of a newly founded daughter company. In general, a loan backed up by a collateral or an external guarantor, should be more secured and get a lower LGD.

Another possible driver of LGD is macro economic variable, that is indicators of how well the economy is performing.[12] Macro variables include growth of GDP, unemployment and interest rates. The rationale is that LGD should be higher when the economic outlook is bad and vice versa. In this thesis, only GDP will be considered since the time period is rather short. Including many variables would increase the risk that one of them would have been well correlated with LGD in the past but with no real relationship. This risk exists for GDP as well, however it is somewhat less since it is a very general economic indicator. Furthermore, by only including one macro economic variable, the risk of nonsense correlation is reduced, and at the same time, a large portion of the real explanatory power is kept since different macro economic variables tend to be correlated with each other.

In addition to these known drivers, several other variables that describe a customer will be included. For these, it is more uncertain whether they have an effect on LGD levels, and in what direction if any.

[10] Hlawatsch and Ostrowski, 'Simulation and estimation of loss given default'; Johnston Ross and Shibut, 'What Drives Loss Given Default? Evidence from Commercial Real Estate Loans at Failed Banks'.

[11] Frontczak and Rostek, 'Modeling loss given default with stochastic collateral'.

[12] Bellotti and Crook, 'Loss given default models incorporating macroeconomic variables for credit cards'.

Table 1: *Explanatory variables to be used in the modelling, divided into idiosyncratic variables, macro economic variables, and time.*

Securitisation Ratio Real Estate	*Share of exposure being secured by real estate collaterals*
Securitisation Ratio Guarantees	*Share of exposure being secured by an external guarantor*
Securitisation Ratio Other	*Share of exposure being secured by other types of collaterals*
Fully secured dummy	*Indicator of whether the customer is 100% secured by collaterals*
SME dummy	*Indicator of whether the customer is a smaller or larger company*
Exposure at Default	*The (logged) exposure value at default*
GDP growth	*Monthly, seasonally adjusted GDP growth in Sweden*
Time	*Number of months that has elapsed since December 2007*

Since the interest lays in loss given default, only defaulted customers are included. The data that will be used for the modelling come from Nordea's corporate Swedish portfolio, and consists of realised LGDs, explanatory variables and data period of default on a business partner level. The time period of the data is 2008 to 2015. Two types of explanatory variables will be used: macro economic variables, which are common for all observations in a given time period, and idiosyncratic variables which are customer specific. The explanatory variables that will be used are listed in Table 1.

4 Theoretical framework

In this section, the theory behind the approaches that will be used is presented. It will start with standard linear and logistic regression, which will be used in Approach 1. Afterwards, it will present the time varying intercept model and how that can be applied to linear and logistic regression which will be used in Approach 2. Last the theory behind SVM and SVR that applies for Approach 3 and 4 will be presented. In general, when relevant, the algorithms for estimation will be presented together with the models.

4.1 Linear & Logistic Regression

Throughout this section, $\mathbf{y}$ will be the dependent variable, $\boldsymbol{\beta}$ will be the parameter vector and $\mathbf{X}$ will be the design matrix. The notation $\mathbf{x}_i$ refers to the vector of explanatory variables for observation i.

An ordinary linear regression has the form of Equation 4 where ϵ is the error term.

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i \tag{4}$$

A common assumption in a linear regression is that the residuals $\boldsymbol{\epsilon}$ are normally distributed. Whenever this is true, estimates obtained by the method of least squares, see 5, are the maximum likelihood estimates of the parameters.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \sum_i (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \tag{5}$$

When the assumption of normally distributed error does not hold, but the errors still have an expected value of zero and are independent with equal variance, the method of least squares is still the best linear unbiased estimator, according to *Gauss-Markov Theorem*.[13]

4.1.1 Linear regression with latent variables

In a standard setup of linear regression, all values of the explanatory variables are known. In a latent variable model, there are latent i.e. unobservable variables related to the response variable.[14] A general linear regression model with latent variables as regressors has the form of Equation 6 where $\mathbf{w}$ are unknown latent regressors and $\boldsymbol{\gamma}$ their parameters, while $\mathbf{x}$ are standard known regressors.

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \boldsymbol{\gamma}^T \mathbf{w}_i + \epsilon_i \tag{6}$$

[13] Sundberg, *Lineära Statistiska Modeller*, p. 24.
[14] Burnham et al., 'Latent variable multivariate regression modeling'

For estimating a multiple regression model with latent regressors, the distribution of the latent variables should be assumed. In general, some structure is often assumed for $\mathbf{w}_i$, for example that the elements are uncorrelated with each other or $\mathbf{x}_i$, or that $\mathbf{w}_i$ is common for several observations, e.g. all observations during a certain time period.

4.1.2 Logistic regression

A logistic regression is used when the dependent variable is binarily categorical and it assumes that y_i is Bernoulli distributed with probability parameter p_i equal to $logit(\boldsymbol{\beta}^T\mathbf{x}_i)$.[15] There is no closed form solution for the maximum-likelihood estimates of $\boldsymbol{\beta}$, so numerical techniques like Newton-Raphson's method are needed. Since they are very standard and included in most statistical software packages no further description will be given.

4.2 Time varying intercept model

This model is a self designed model in order to explicitly model the time dependency with cross-sectional data with a time dimension.[16] The presentation will first be in the general *Generalised Linear Model* (GLM) form, which is a broad class of models that includes both linear and logistic regression. In a GLM model, the expected value is equal to $g^{-1}(\boldsymbol{\beta^T}\mathbf{x})$ for some invertible function g, referred to as the *link function*, which differs between models in GLM. For the purposes of this thesis, GLM will only be used in the start of this section, since it allows for a very general presentation of the time varying intercept model. When the model has been presented, it will be described how it is used in the case of linear and logistic regression.

The time varying intercept model assumes that the intercept is a random variable that moves like a random walk over time. The distribution of the response variable is assumed to be a GLM when conditioning on the intercept. The intercept for time t is denoted α_t and assumed to vary between constant time periods according to a normal distribution with variable drift but constant volatility. The drift is assumed to be linear in some known regressors $\mathbf{z}$. These are referred to as *time regressors* and are explanatory variables which have a value for each time period. In this thesis, macro economic variables, like GDP, will be used as time regressors. The parameter vector $\boldsymbol{\theta}$ forms a linear relationship between the time regressors and the change in α over two consecutive time periods. This makes the intercept vary according to Equation 7.

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1} + \boldsymbol{\theta}^T\mathbf{z}_t, \sigma_\alpha^2) \tag{7}$$

[15] The *logit*-function is defined as $logit(\cdot) = \frac{1}{1+e^{-[\cdot]}}$.

[16] Note that the data is not panel data, since every customer is measured at just one time period.

The rationale for the model is that it is a flexible approach when the response varies considerably due to idiosyncratic factors as well as over time. Furthermore, it could yield more reasonable inference possibilities of the significance of the time regressors. When these regressors are included in a linear regression, the significance may be overstated, since several observations at the same time period will make it seem that there are more degrees of freedom than it actually is.

α will from time to time have the subscript i. This should be interpreted as the intercept for the time period that observation i belongs to. Remark that if two observations i and j belong to the same time period, then $\alpha_i = \alpha_j$. For brevity, there will be no index t for y since there is a surjective mapping from i to t.

Throughout the thesis, it will be assumed that there are T time periods, referred to with the subscript t, numbered from 1 to T. It will also be assumed that there are N observations in total, which in general will be referred to with the subscript i. N_t will denote the number of observations in time period t. In general, $\sum_t$ should be read as the sum from $t = 1$ to T, except when a variable, e.g. α, is included with subscript $t-1$, then it should be read as the sum from $t = 2$ to T. $\sum_i$ should be read as the sum over all observations and $\sum_{i \in t}$ as the sum over all observations in time period t.

Conditioned on the intercept, the expected value for Y_i is 8.

$$E(Y_i|\alpha_i) = g^{-1}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x_i} + \alpha_i) \tag{8}$$

Denote the distribution function for observation i in the sample as f_{Y_i}. The distribution for the whole sample conditioned on $\boldsymbol{\alpha}$ is then 9.

$$f_{\mathbf{Y}|\boldsymbol{\alpha}} = \prod_i f_{Y_i|\boldsymbol{\alpha}} \tag{9}$$

The joint distribution of $\mathbf{Y}$ and $\boldsymbol{\alpha}$ is expanded in 10.

$$f_{\mathbf{Y},\boldsymbol{\alpha}} = f_{\mathbf{Y}|\boldsymbol{\alpha}} \cdot f_{\boldsymbol{\alpha}} = (\prod_i f_{Y_i|\boldsymbol{\alpha}}) f_{\boldsymbol{\alpha}} \tag{10}$$

The distribution function of $\boldsymbol{\alpha}$ is expanded in 11 where F_t denotes the time sequence of $\boldsymbol{\alpha}$ up until time period t and T is the latest time period.

$$f_{\boldsymbol{\alpha}} = f_{\alpha_T|F_{T-1}} \cdot f_{\alpha_{T-1}|F_{T-2}} \cdot \cdots \cdot f_{\alpha_2|F_1} \cdot f_{\alpha_1} \tag{11}$$

The conditional distributions in 11 are easily obtained since $\alpha_t|F_{t-1} \sim N(\alpha_{t-1} + \boldsymbol{\theta}^T z_t, \sigma_\alpha^2)$. The unconditional distribution of α_1 remains to be specified. Instead, α_1 is conditioned on being 0. This will imply no restriction, since it means that the ordinary intercept contained within the design matrix $\mathbf{X}$ states the base level, that is the intercept at time period 1.

The joint probability density function of $\boldsymbol{\alpha}$ is thus proportional to 12.

$$f_{\boldsymbol{\alpha}} \propto \sigma_{\alpha}^{-T} e^{-\frac{1}{2}\sigma_{\alpha}^{-2}\sum_t(\alpha_t - \alpha_{t-1} - \boldsymbol{\theta}^T \mathbf{z}_i)^2} \tag{12}$$

This makes the joint probability density function of the sample $\mathbf{Y}$ and $\boldsymbol{\alpha}$ proportional to 13.

$$f_{\mathbf{Y},\boldsymbol{\alpha}} = f_{\mathbf{Y}|\boldsymbol{\alpha}} \cdot f_{\boldsymbol{\alpha}} \propto f_{\mathbf{Y}|\boldsymbol{\alpha}} \cdot \sigma_{\alpha}^{-T} e^{-\frac{1}{2}\sigma_{\alpha}^{-2}\sum_t(\alpha_t - \alpha_{t-1} - \boldsymbol{\theta}^T \mathbf{z}_i)^2} \tag{13}$$

4.2.1 Expectation-Maximisation algorithm

The standard analytical method for maximum likelihood estimation when latent variables are present is to integrate over the latent variables which would yield the distribution of $\mathbf{Y}$ without $\boldsymbol{\alpha}$ being present. However in 13, both parts contain elements of $\boldsymbol{\alpha}$, possible in very complicated ways depending on $f_{\mathbf{Y}|\boldsymbol{\alpha}}$. In general, this will make it intractable to integrate away $\boldsymbol{\alpha}$, making the way for numerical techniques. *Expectation-Maximisation* algorithm is a technique of that nature. It is an iterative method to find the maximum likelihood estimates when latent variables are presented. It is guaranteed to converge to a local maximum, though not global.[17]

EXPECTATION MAXIMISATION ALGORITHM

1. Guess some start values of the parameters
2. Repeat until convergence
   (a) Calculate the distribution of the latent variables given current parameter values
   (b) Calculate the expected value of the log likelihood given the current estimate distribution of the latent variables
   (c) Find the new parameters by maximising the expected log likelihood from step (b)

4.2.2 Linear regression with time varying intercept

Now to applying the time varying intercept for a linear regression. It will be assumed that the elements in $\mathbf{Y}$ conditioned on $\boldsymbol{\alpha}$ are independent and normally distributed with mean vector equal to $\mathbf{X} \cdot \boldsymbol{\beta}$ and variance equal to σ_{ϵ}^2. This means that conditioned on $\boldsymbol{\alpha}$, the inference problem becomes a standard linear regression.

[17] Sundberg, *Statistical Modelling by Exponential Families*, p. 113.

Combining the conditional distribution of $\mathbf{y}$ and 13 makes the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\alpha}$ become proportional to 14. As before, N denotes the number of observations in total, and T denotes the number of time periods in total.

$$f_{\mathbf{Y},\boldsymbol{\alpha}} \propto \sigma_\epsilon^{-N}\sigma_\alpha^{-T}e^{-\frac{1}{2}\sigma_\epsilon^{-2}\sum_i(y_i-\alpha_i-\boldsymbol{\beta}^T\mathbf{x}_i)^2}e^{-\frac{1}{2}\sigma_\alpha^{-2}\sum_t(\alpha_t-\alpha_{t-1}-\boldsymbol{\theta}^T\mathbf{z}_i)^2} \tag{14}$$

From Equation 14, the log-likelihood is stated in Equation 15.

$$\begin{aligned} &-\frac{1}{2}\frac{\sum_i(y_i-\alpha_t-\boldsymbol{\beta}^T\mathbf{x}_i)^2}{\sigma_\epsilon^2} - N\cdot log(\sigma_\epsilon) \\ &-\frac{1}{2}\frac{\sum_t(\alpha_t-\alpha_{t-1}-\boldsymbol{\theta}^T\mathbf{z}_t)^2}{\sigma_\alpha^2} - T\cdot log(\sigma_\alpha) \end{aligned} \tag{15}$$

Since the elements in $\mathbf{y}$ and $\boldsymbol{\alpha}$ are included only as quadratic and linear terms in Equation 15, it seems like the joint distribution of $\mathbf{y}$ and $\boldsymbol{\alpha}$ is normal. Proving this would be equivalent to showing that the covariance matrix is positive semidefinitive. In general, the distribution of some of the random variables in a normal distribution may be found merely by omitting the row in the mean vector, and the row and column in the covariance matrix. However, in practise, it would be very hard to do this. The covariance matrix is obtained by first calculating the precision matrix C^{-1}, whose elements are found by the second derivative with respect to α and y. This will be dependent on many parameters and intractable to invert.[18] Therefore it will not be investigated further whether the joint distribution of $\mathbf{y}$ and $\boldsymbol{\alpha}$ is normally distributed.

Using Expectation-Maximisation algorithm for time varying intercept in linear regression

Since an analytical solution was intractable, the Expectation-Maximisation algorithm will be used in order to obtain the maximum likelihoood estimates of $\boldsymbol{\beta}$. The rest of the section will be about using this algorithm in order to calibrate the model. The start values chosen for the parameters are the ones obtained from a regression without a time varying intercept, i.e. an ordinary linear regression. This section will first describe how to calculate the distribution of $\boldsymbol{\alpha}$ given current parameters, and then proceed with how to update the parameters given the new distribution of $\boldsymbol{\alpha}$.

Calculate the conditional distribution of $\boldsymbol{\alpha}$ given current parameters

By conditioning Equation 15 on $\mathbf{y}$, the distribution function of $\boldsymbol{\alpha}$ becomes proportional to Equation 16, where $\mathbf{y}$ is to be considered as fixed.

$$-\frac{1}{2}\frac{\sum_t\alpha_t^2+\alpha_{t-1}^2-2\alpha_t\alpha_{t-1}-2(\alpha_t-\alpha_{t-1})\hat{\boldsymbol{\theta}}^T\mathbf{z}_t}{\hat{\sigma}_\alpha^2} - \frac{1}{2}\frac{\sum_i\alpha_i^2+2\alpha_i(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i-y_i)}{\hat{\sigma}_\epsilon^2}$$

[18] The precision matrix is the inverse of the covariance matrix.

(16)

Let N_t as before be the number of observations at time t. Then Equation 16 may be written as Equation 17.

$$
\begin{aligned}
&-\frac{1}{2}\sum_t \alpha_t^2\Big(\frac{1_{t\neq 1}+1_{t\neq T}}{\hat{\sigma}_\alpha^2}+\frac{N_t}{\hat{\sigma}_\epsilon^2}\Big)\\
&-\frac{1}{2}\sum_t \frac{\alpha_t\alpha_{t-1}}{\hat{\sigma}_\alpha^2}\\
&-\frac{1}{2}\sum_t 2\alpha_t\Big(\frac{\sum_{i\in t}\hat{\boldsymbol{\beta}}^T\mathbf{x}_i-y_i}{\hat{\sigma}_\alpha^2}+1_{t\neq 1}\cdot\hat{\boldsymbol{\theta}}^T\mathbf{z}_{t-1}+1_{t\neq T}\cdot\hat{\boldsymbol{\theta}}^T\mathbf{z}_t\Big) \qquad (17)
\end{aligned}
$$

Since the log-likelihood of a normal distribution is proportional to $-\frac{1}{2}\sum_{ij}(x_ix_j - x_i\mu_j)\cdot C_{ij}^{-1}$, the likelihood in Equation 17 can be assessed to be normally distributed if it can be shown that the precision matrix C is positive semi-definite. First, C^{-1} is explicitly calculated in Equation 18 by Equation 17 where $\mathcal{L}$ denotes the likelihood function.

$$
C_{ij}^{-1}=-\frac{1}{2}\frac{\partial^2\mathcal{L}}{\partial\alpha_i\partial\alpha_j}=\begin{cases}(1_{t\neq 1}+1_{t\neq T})\hat{\sigma}_\alpha^{-2}+N_t\hat{\sigma}_\epsilon^{-2} & for \quad i=j\\ \hat{\sigma}_\alpha^{-2} & for \quad i=j\pm 1\\ 0 & \text{otherwise}\end{cases} \qquad (18)
$$

The precision matrix, $\mathbf{C}^{-1}$, is displayed in full matrix format in 19.

$$
2\cdot\begin{pmatrix}
(\hat{\sigma}_\alpha^{-2}+N_t\hat{\sigma}_\epsilon^{-2}) & \hat{\sigma}_\alpha^{-2} & 0 & \dots & 0 & 0\\
\hat{\sigma}_\alpha^{-2} & (2\hat{\sigma}_\alpha^{-2}+N_t\hat{\sigma}_\epsilon^{-2}) & \hat{\sigma}_\alpha^{-2} & \dots & 0 & 0\\
0 & \hat{\sigma}_\alpha^{-2} & (2\hat{\sigma}_\alpha^{-2}+N_t\hat{\sigma}_\epsilon^{-2}) & \ddots & 0 & 0\\
\vdots & \vdots & \ddots & \ddots & \hat{\sigma}_\alpha^{-2} & 0\\
0 & 0 & 0 & \hat{\sigma}_\alpha^{-2} & (2\hat{\sigma}_\alpha^{-2}+N_t\hat{\sigma}_\epsilon^{-2}) & \hat{\sigma}_\alpha^{-2}\\
0 & 0 & 0 & 0 & \hat{\sigma}_\alpha^{-2} & (\hat{\sigma}_\alpha^{-2}+N_t\hat{\sigma}_\epsilon^{-2})
\end{pmatrix} \qquad (19)
$$

A matrix $\mathbf{A}$ is positive definite if and only if, for all vectors $\mathbf{s}$, it holds that $\mathbf{s}^T\mathbf{A}\mathbf{s}\geq 0$. Standard results from linear algebra says that a matrix is positive semi-definite if and only if its inverse is. It is thus sufficient to show that $\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}\geq 0$ for the covariance matrix to be positive semi-definite and the

distribution of $\boldsymbol{\alpha}$ to be normal. This is shown in Equation 20.

$$\begin{aligned}
2 \cdot \mathbf{s}^T \mathbf{C}^{-1} \mathbf{s} &= \sum_i s_i^2 (\hat{\sigma}_\alpha^{-2}(1_{t\neq 1} + 1_{t\neq T}) + N_s \hat{\sigma}_\epsilon^{-2}) + 2 \sum_{i=j+1} s_i s_j \hat{\sigma}_\alpha^{-2} \\
&\geq \sum_i s_i^2 (\hat{\sigma}_\alpha^{-2}(1_{t\neq 1} + 1_{t\neq T}) + 2 \sum_{i=j+1} s_i s_j \hat{\sigma}_\alpha^{-2} \\
&= \hat{\sigma}_\alpha^{-2} \Big(\sum_i s_i^2 (1_{t\neq 1} + 1_{t\neq T}) + 2 \sum_{i=j+1} s_i s_j \Big) \\
&= \hat{\sigma}_\alpha^{-2} \Big(\sum_i (s_i + s_{i+1})^2 \Big) \\
&\geq 0
\end{aligned} \tag{20}$$

This means that the likelihood is normal. In theory the mean vector, $\boldsymbol{\mu}$, could be analytically solved for, like the precision matrix, but it would give a very complicated expression not tenable for estimation. Instead, it is solved by optimising one α_t at a time while holding the other constant. Since the log-likelihood of a normal distribution is concave, this will yield the mode value of the function, which will coincide with the expected value. By differentiating Equation 15 with respect to α_t, setting the derivative to zero, and solving for α_t, the optimum value is found. Standard differentiation and algebra yields Equation 21.

$$\alpha_t = \frac{1_{t\neq 1} \cdot \dfrac{\alpha_{t-1} + \hat{\boldsymbol{\theta}}^T \mathbf{z}_t}{\hat{\sigma}_\alpha^2} + 1_{t\neq T} \cdot \dfrac{\alpha_{t+1} - \hat{\boldsymbol{\theta}}^T \mathbf{z}_{t+1}}{\hat{\sigma}_\alpha^2} + \dfrac{\sum_i y_i - \hat{\beta} x_i}{\hat{\sigma}_\epsilon^2}}{\dfrac{1_{t\neq 1} + 1_{t\neq N_t}}{\hat{\sigma}_\alpha^2} + \dfrac{N_t}{\hat{\sigma}_\epsilon^2}} \tag{21}$$

By iterating Equation 21 over all elements in $\boldsymbol{\alpha}$, the mean vector will be found. The mean vector in combination with the precision matrix completely characterises the normal distribution which will be used for the expectation step.

Taking expectation of likelihood for obtaining parameter estimates

Now when the distribution of $\boldsymbol{\alpha}$ given the current estimates has been obtained, it is time for the expectation step in the EM algorithm. Taking the expectation of 15 yields 22.

$$E\Big(-T \cdot log(\sigma_\alpha) - N \cdot log(\sigma_\epsilon) - \frac{1}{2} \frac{\sum_i (y_i - \boldsymbol{\beta}^T \mathbf{x}_i - \alpha_t)^2}{\sigma_\epsilon^2} - \frac{1}{2} \frac{\sum_t (\alpha_t - \alpha_{t-1} - \boldsymbol{\theta}^T \mathbf{z}_i)^2}{\sigma_\alpha^2} \Big)$$

$$
\begin{aligned}
&= -T{\cdot}log(\sigma_\alpha) - N{\cdot}log(\sigma_\epsilon) - \frac{1}{2}\frac{\sum_i (y_i - \boldsymbol{\beta}^T\mathbf{x}_i)^2 + E(\alpha_t^2) - 2E(\alpha_t)(y_i - \boldsymbol{\beta}^T\mathbf{x}_i)}{\sigma_\epsilon^2} \\
&- \frac{1}{2}\frac{\sum_t E(\alpha_t^2) + E(\alpha_{t-1}^2) - 2\cdot E(\alpha_{t-1}\alpha_t) + (\boldsymbol{\theta}^T\mathbf{z}_t)^2 + 2\boldsymbol{\theta}^T\mathbf{z}_i(E(\alpha_{t-1}) - E(\alpha_t))}{\sigma_\alpha^2} \\
&= -T{\cdot}log(\sigma_\alpha) - N{\cdot}log(\sigma_\epsilon) - \frac{1}{2}\frac{\sum_i (y_i - \boldsymbol{\beta}^T\mathbf{x}_i - E(\alpha_t))^2 + \overbrace{E(\alpha_t^2) - E(\alpha_t)^2}^{=Var(\alpha_t)}}{\sigma_\epsilon^2} \\
&- \frac{1}{2}\frac{\sum_t (E(\alpha_{t-1}) - E(\alpha_t) + \boldsymbol{\theta}^T\mathbf{z}_t)^2 + \overbrace{E(\alpha_t^2) + E(\alpha_{t-1}^2) - 2\cdot E(\alpha_{t-1}\alpha_t) - (E(\alpha_{t-1}) - E(\alpha_t))^2}^{=Var(\alpha_t - \alpha_{t-1})}}{\sigma_\alpha^2} \\
&= -T\cdot log(\sigma_\alpha) - N\cdot log(\sigma_\epsilon) - \frac{1}{2}\frac{\sum_i (y_i - \boldsymbol{\beta}^T\mathbf{x}_i - E(\alpha_t))^2 + Var(\alpha_i)}{\sigma_\epsilon^2} \\
&- \frac{1}{2}\frac{\sum_t (E(\alpha_{t-1}) - E(\alpha_t) + \boldsymbol{\theta}^T\mathbf{z}_t)^2 + Var(\alpha_t - \alpha_{t-1})}{\sigma_\alpha^2}
\end{aligned}
\tag{22}
$$

Note that Equation 22 corresponds to two linear regressions, one with $\mathbf{y}$ as response, $\mathbf{x}$ as regressors, $E(\boldsymbol{\alpha})$ as offset, and $\boldsymbol{\beta}$ as parameters, and another one with $E(\alpha_{t-1}) - E(\alpha_t)$ as response, $\mathbf{z}$ as regressors, and $\boldsymbol{\theta}$ as parameters. This means that the estimates for β and θ may be obtained from ordinary least squares obtained by Equation 23 and Equation 24.

$$
\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \sum_i (y_i - \boldsymbol{\beta}^T\mathbf{x}_i - E(\alpha_i))^2 \tag{23}
$$

$$
\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{argmin} \sum_t (E(\alpha_t) - E(\alpha_{t-1}) - \boldsymbol{\theta}^T\mathbf{z}_t)^2 \tag{24}
$$

The variance estimators will change some compared to the standard estimates due to the randomness of the time varying intercept. By differentiating Equation 22 with respect to σ_α and σ_ϵ and setting the derivative to zero yields Equation 25 and 26 are obtained.

$$
\hat{\sigma}_\epsilon^2 = \frac{\sum_i (y_i - \boldsymbol{\beta}^T\mathbf{x}_i - E(\alpha_t))^2 + Var(\alpha_i)}{N} \tag{25}
$$

$$
\hat{\sigma}_\alpha^2 = \frac{\sum_t (E(\alpha_{t-1}) - E(\alpha_t) + \boldsymbol{\theta}^T\mathbf{z}_t)^2 + Var(\alpha_t - \alpha_{t-1})}{T} \tag{26}
$$

SUMMARY OF LINEAR REGRESSION WITH TIME VARYING INTERCEPT

1. Obtain start values of parameters by doing an ordinary linear regression
2. Repeat until convergence
   (a) Calculate the distribution of $\boldsymbol{\alpha}$ by Equation 19 and Equation 21
   (b) Update the parameters according to Equation 23-26

4.2.3 Logistic regression with time varying intercept

Now a logistic regression with time varying intercept will be constructed. Then the likelihood conditioned on the time varying intercept will be Bernoulli distributed. Let $\mathbf{y}$ be a set of outcomes of Bernoulli variables, with probability parameter $p_i = logit(\mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\alpha}_i)$. Denote $q_i = 1 - p_i$. Then the distribution for $\mathbf{y}$, using Equation 13, is proportional Equation 27.

$$f_{\mathbf{Y}} = (\prod_i {p_i}^{y_i} {q_i}^{1-y_i}) \cdot \sigma_\alpha^{-2} e^{-\frac{1}{2}\sigma_\alpha^{-2}\sum_t(\alpha_t - \alpha_{t-1} - \boldsymbol{\theta}^T\mathbf{z}_i)^2} \tag{27}$$

The log-likelihood of $\mathbf{y}$ then becomes proportional to Equation 28.

$$\sum_i y_i \cdot log(p_i) + (1 - y_i) \cdot log(q_i)$$
$$-\frac{1}{2}\sum_t \frac{(\alpha_t - \alpha_{t-1} - \hat{\boldsymbol{\theta}}^T z_t)^2}{\sigma_\alpha^2} - T \cdot log(\sigma_\alpha) \tag{28}$$

The problem in Equation 28 is very hard to solve, even when using the Expectation-Maximisation algorithm, due to the logit-expressions contained in p_i and q_i. Instead, a second-degree Taylor expansions that will be done around the current estimated expected value of α will replace the logit-expression. This will transform the problem to a quadratic one, which can be solved with the same methods as for the linear regression. Approximating a complicated likelihood with a quadratic function is a known method which also yields asymptotically the same maximum likelihood estimates as the real one.[19]

Approximating the likelihood with Taylor expansions

Denote the current expected value of $\boldsymbol{\alpha}$ as $\hat{\boldsymbol{\mu}}$, i.e. $E_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}}(\boldsymbol{\alpha}) = \hat{\boldsymbol{\mu}}$. Then the log-likelihood in Equation 28 can be approximated by two Taylor expansions, for $logit(\pm(\alpha_t + \hat{\boldsymbol{\beta}}^T\mathbf{x}_i))$. The first Taylor expansion is calculated in Equa-

[19] Held and Sabanés Bové, *Applied Statistical Inference*, p 34-35.

tion 29.

$$
\begin{aligned}
& log\Big(\frac{1}{1+e^{-(\alpha_t+\hat{\boldsymbol{\beta}}^T\mathbf{x}_i)}}\Big) \\
\approx\ & log(\frac{1}{1+e^{-(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t)}}) + \frac{(\alpha_t-\hat{\mu}_t)}{1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}} - \frac{(\alpha_t-\hat{\mu}_t)^2}{2}\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2} \\
\propto\ & \frac{\alpha_t}{1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}} - \frac{(\alpha_t-\hat{\mu}_t)^2}{2}\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2} \\
\propto\ & \alpha_t\cdot\Big(\frac{1}{1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}} + \hat{\mu}_t\cdot\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2}\Big) - \frac{\alpha_t^2}{2}\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2}
\end{aligned}
\tag{29}
$$

The second Taylor expansion is calculated in Equation 30.

$$
\begin{aligned}
& log\Big(\frac{1}{1+e^{\alpha_t+\hat{\boldsymbol{\beta}}^T\mathbf{x}_i}}\Big) \\
\approx\ & log(\frac{1}{1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}) - \frac{(\alpha_t-\hat{\mu}_t)}{1+e^{-(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t)}} - \frac{(\alpha_t-\hat{\mu}_t)^2}{2}\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2} \\
\propto\ & -\frac{\alpha_t}{1+e^{-(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t)}} - \frac{(\alpha_t-\hat{\mu}_t)^2}{2}\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2} \\
\propto\ & \alpha_t\cdot\Big(-\frac{1}{1+e^{-(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t)}} + \hat{\mu}_t\cdot\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2}\Big) - \frac{\alpha_t^2}{2}\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2}
\end{aligned}
\tag{30}
$$

Inserting Equation 29 and Equation 30 into the Bernoulli part, i.e. ${p_i}^{y_i}{q_i}^{1-y_i}$ of Equation 28 yields Equation 31.

$$
\begin{aligned}
& \sum_i y_i\cdot log(p_i) + (1-y_i)\cdot log(q_i) \\
=& \sum_i y_i\cdot log\Big(\frac{1}{1+exp(-\alpha_t-\hat{\boldsymbol{\beta}}^T\mathbf{x}_i)}\Big) + (1-y_i)\cdot log\Big(\frac{1}{1+exp(\alpha_t+\hat{\boldsymbol{\beta}}^T\mathbf{x}_i)}\Big) \\
\underset{\sim}{\propto}& \sum_i y_i\cdot\Big(\alpha_t\cdot\Big(\frac{1}{1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}} + \hat{\mu}_t\cdot\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2}\Big) - \frac{\alpha_t^2}{2}\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2}\Big) \\
& + (1-y_i)\cdot\Big(\alpha_t\cdot\Big(-\frac{1}{1+exp(-(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t))} + \hat{\mu}_t\cdot\frac{exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t)}{(1+exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t))^2}\Big) - \frac{\alpha_t^2}{2}\frac{exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t)}{(1+exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t))^2}\Big) \\
=& \sum_i \alpha_t^2(-\frac{1}{2}\underbrace{\frac{e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t}}{(1+e^{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t})^2}}_{=k_i}) + \alpha_t\underbrace{\Big(\hat{\mu}_t\cdot\frac{exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t)}{(1+exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_i+\hat{\mu}_t))^2} + \frac{1}{1+exp(\hat{\mu}_t+\hat{\boldsymbol{\beta}}^T\mathbf{x}_i)} + y_i - 1\Big)}_{=c_i}
\end{aligned}
\tag{31}
$$

The log-likelihood from Equation 28 may therefore be approximated by Equation 32.

$$
\begin{aligned}
&- T \cdot log(\sigma_\alpha) - \frac{1}{2} \sum_t \frac{(\alpha_t - \alpha_{t-1} - \hat{\boldsymbol{\theta}}^T z_t)^2}{\sigma_\alpha^2} \\
&+ \sum_i \alpha_t c_i - \frac{1}{2} \alpha_t^2 k_i
\end{aligned} \tag{32}
$$

Calculate the conditional distribution of $\boldsymbol{\alpha}$ given current parameters

Note that Equation 32 is the log-likelihood of a normal distribution. The mode vector of $\boldsymbol{\alpha}$ then corresponds to the mean vector, which as in the case with linear regression is found by optimising the different α_t one at a time while holding the other ones constant. This is done by differentiating Equation 32 with respect to α_t, setting the derivative to zero and solving for α_t, which yields Equation 33 after standard, but tedious algebra.

$$
\alpha_t = \frac{1_{t \neq 1} \cdot \frac{\alpha_{t-1} + \hat{\boldsymbol{\theta}}^T z_t}{\hat{\sigma}_\alpha^2} + 1_{t \neq T} \cdot \frac{\alpha_{t+1} - \hat{\boldsymbol{\theta}}^T z_{t+1}}{\hat{\sigma}_\alpha^2} + \sum_{i \in t} c_i \Big)}{\frac{1_{t \neq 1} + 1_{t \neq T}}{\hat{\sigma}_\alpha^2} + \sum_{i \in t} k_i} \tag{33}
$$

As with the linear regression the elements in the precision matrix can be calculated by differentiating the likelihood twice with respect to $\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j$. C^{-1} is explicitly calculated in Equation 34 by differentiating the likelihood in Equation 32. As before $\mathcal{L}$ denotes the likelihood function.

$$
C_{ij}^{-1} = -\frac{1}{2} \frac{\partial^2 \mathcal{L}}{\partial \alpha_i \partial \alpha_j} = \begin{cases} (1_{t \neq 1} + 1_{t \neq T}) \hat{\sigma}_\alpha^{-2} + 2 \sum_{i \in t} k_i & i = j \\ \hat{\sigma}_\alpha^{-2} & i = j \pm 1 \\ 0 & \text{otherwise} \end{cases} \tag{34}
$$

Taking expectation of likelihood for obtaining parameter estimates

When the distribution of $\boldsymbol{\alpha}$ has been found, the expectation step will make use of Gauss approximation formula for optimising the parameters, see Equation 35.[20]

$$
\begin{aligned}
E(log(f_{\mathbf{Y},\boldsymbol{\alpha}})) \propto E\Bigg(\sum_i y_i \cdot log\Big(\frac{1}{1 + exp(-\alpha_t - \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}\Big) + (1 - y_i) \cdot log\Big(\frac{1}{1 + exp(\alpha_t + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}\Big) \Bigg) \\
+ E\Bigg(- 1_{t \neq 1} \frac{1}{2} \cdot \frac{(\alpha_t - \alpha_{t-1} - \hat{\boldsymbol{\theta}}^T z_t)^2}{\sigma_\alpha^2} - 1_{t \neq T} \frac{1}{2} \cdot \frac{(\alpha_{t+1} - \alpha_t - \hat{\boldsymbol{\theta}}^T z_{t+1})^2}{\sigma_\alpha^2} \Bigg)
\end{aligned} \tag{35}
$$

[20] Gauss approximation formula is just a first order Taylor expansion employed when calculating the expected value of a random variable, i.e. $E(g(X)) \approx g(E(X))$.

The second part of Equation 35 is identical to the corresponding part in the linear regression, and the estimates for $\boldsymbol{\theta}$ and σ_α^2 are therefore the same in the logistic case as in the linear regression. For the first part of Equation 35, Gauss approximation formula is utilised, see Equation 36.

$$\begin{aligned} & E\Bigg(\sum_i y_i \cdot log\Big(\frac{1}{1+exp(-\alpha_t - \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}\Big) + (1-y_i) \cdot log\Big(\frac{1}{1+exp(\alpha_t + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}\Big)\Bigg) \\ \approx & \sum_i y_i \cdot log\Big(\frac{1}{1+exp(-E(\alpha_t) - \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}\Big) + (1-y_i) \cdot log\Big(\frac{1}{1+exp(E(\alpha_t) + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}\Big) \end{aligned} \tag{36}$$

Equation 36 means that $\boldsymbol{\beta}$ may be solved for by using standard logistic regression with $E(\boldsymbol{\alpha})$ as an offset. This procedure of optimising $\boldsymbol{\alpha}$ and then the parameters will be repeated until convergence.

SUMMARY OF LOGISTIC REGRESSION WITH TIME VARYING INTERCEPT

1. Obtain start values of parameters by doing an ordinary logistic regression
2. Repeat until convergence
   (a) Calculate the distribution of $\boldsymbol{\alpha}$ by Equation 33 and Equation 34
   (b) Update the $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}_\alpha^2$ according to Equation 24 and 26.
   (c) Update $\hat{\boldsymbol{\beta}}$ by a logistic regression with $E(\boldsymbol{\alpha})$ an offest variable

4.3 Support Vector Machines

Machine learning is an academic field in the intersection between statistics and computer science. The line between statistics and machine learning is fine and differs between authors. Like statistics, machine learning deals with analysis and interpretation of data, but has more focus on prediction and less on inference. Furthermore, while methods in statistics often tend to have at least some distributional assumptions and make use of mathematics to draw asymptotic conclusions which are the the primary method justifications, machine learning is almost always concerned with large data, focusing less on assumptational justifications and instead of evaluating the usefulness of models by *cross validation*, i.e. testing the models on data not used for development.

Support Vector Machine (SVM) is a method within machine learning which in its original form was developed in the 60s under the name *Generalized Portrait* algorithm. In the 90s it took its modern form by including mapping to higher dimensional space and a soft margin formulation.[21] SVM has become very popular due to its flexibility of the problem statement, the relatively easy-to-use due to its foundation of convex optimisation, the both theoretical and intuitive justification and the empirical good performance. They have historically been computationally expensive and still are for very large datasets, so the development of more powerful computers have made the method feasible for larger and larger data.

A SVM is calibrated by a dataset $\{d_i, \mathbf{x}_i\}_{i=1}^N$ where $\{d_i, \mathbf{x}_i\} \in \{\{-1, 1\}, \mathbb{R}^n\}$ with d being the label and $\mathbf{x} \in \mathbb{R}^n$, which may be discrete as well as continuous, is the *feature vector*, which is the machine learning equivalent of explanatory variables. After calibration, the purpose of the model is to with as good accuracy as possible predict the new label given a new feature vector.

The original and most straightforward formulation of SVMs operate in the input space of the features, and applies only to problem with linearly separable labels. Two generalisations have been vital to its usefulness: the first being the mapping of the features to higher dimensional space, where the labels are more easily separable, and the second being the soft margin formulation which allows some misclassified points in the training data and instead impose a penalty on misclassifications in the optimisation problem.

The outline of the section is the following: (i) a presentation of the problem in the linearly separable case in input space, (ii) a generalisation to allow for mapping to higher dimensional space, (iii) a generalisation to the non-linearly separable case, (iv) a solution of the SVM problem in higher dimensions in the non-linearly separable case and lastly (v) a method for obtaining probabilities from classifications.

[21] Smola and Schölkopf, 'A tutorial on support vector regression'.

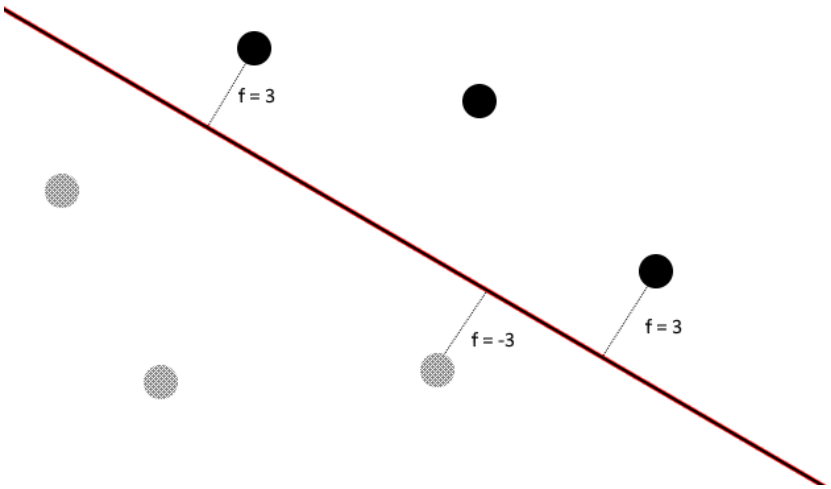

Figure 1: *The figure shows points with two different labels, black and grey in* $\mathbb{R}^2$*. The red line separates the points with the greatest possible margin, which is 3 in this case.* f *is the discriminant function which measures a signed distance to the hyperplane.*

4.3.1 Linearly separable case in input space

A general hyperplane in $\mathbb{R}^n$ is characterised by a vector $\mathbf{w}$ and a constant b and is built up of all points $\mathbf{x} \in \mathbb{R}^n$ that satisfies Equation 37. A standard result from linear algebra also says that $\mathbf{w}$ is a normal vector to the hyperplane.

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{37}$$

The objective of an SVM in the linearly separable case with features $\mathbf{x}_i \in \mathbb{R}^n$ is to find the hyperplane in $\mathbb{R}^n$ which separates the points with labels, referred to as d, with the values 1 and -1 with the widest margin. This is done by maximisation of the minimum distance between the points and the hyperplane under the constraint that the points with different labels are separated into different side on the hyperplane, which can be seen in Figure 1. The vector $\mathbf{w}$, referred to as *weights*[22] and the constant b, referred to as *bias*[23], build up the hyperplane. Let r_i be the signed distance from the hyperplane to the observation i. The SVM can then be stated as Maximisation Problem 38.

$$\begin{aligned} \max_{\mathbf{w},b} \quad & min(\mathbf{r}) \\ & s.t.\ \mathbf{w}^T\mathbf{x_i} + b \geq 0 \qquad \forall i : d_i = +1 \\ & s.t.\ \mathbf{w}^T\mathbf{x_i} + b < 0 \qquad \forall i : d_i = -1 \end{aligned} \tag{38}$$

[22]The corresponding name for *parameters* in machine learning
[23]The corresponding name for *intercept* in machine learning

Since Maximisation Problem 38 does not bring an obvious solution, it will be reformulated in several steps to equivalent optimisation problems. In a first step, the weights and the bias are rescaled so the constraints change to those in 39.

$$\begin{aligned} \max_{\mathbf{w},b} \quad & min(\mathbf{r}) \\ & \mathbf{w}^T\mathbf{x_i} + b \geq +1 \qquad \forall i : d_i = +1 \\ & \mathbf{w}^T\mathbf{x_i} + b \leq -1 \qquad \forall i : d_i = -1 \end{aligned} \tag{39}$$

In the next step, the so called *discriminant function* is introduced, see Definition 40.

$$f(\mathbf{x}_i) = f_i = \mathbf{w}^T\mathbf{x}_i + b \tag{40}$$

The discriminant function gives a measure of the signed distance to the hyperplane for a vector $\mathbf{x}$. This can be seen by letting $\mathbf{x}_p$ be the projection of $\mathbf{x}$ on the hyperplane, and r the signed distance from the hyperplane. Since $\frac{\mathbf{w}}{||\mathbf{w}||}$ is the normal vector with unit length, $\mathbf{x}$ can be written as $\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{||\mathbf{w}||}$. This implies that $f = \mathbf{w}^T(\mathbf{x}_p + r\frac{\mathbf{w}}{||\mathbf{w}||}) + b = r||\mathbf{w}||$ since by definition $\mathbf{w}^T\mathbf{x}_p + b = 0$, thus $f_i = r_i||\mathbf{w}||$ which is the signed distance scaled with a positive constant.

When solving 39, it will be assumed that there is at least one point with label 1 that has distance 1 to the hyperplane and at least one point with label -1 that has distance -1 to the hyperplane, i.e. that the inequalities in Maximisation Problem 39 hold as equalities for at least one example each. This is trivial since the distance of the hyperplane would be closer to the points with one of the labels so the minimum distance would not be maximised.

All observations for which it holds that the distance to the hyperplane is ± 1 are referred to as *support vectors*. These are the only vectors that will affect the solution, i.e. when the hyperplane has been found an arbitrarily amount of observations may be added at arbitrarily places as long as the absolute distance to the hyperplane is one or greater. This property of support vector is so important that it has given the name for the model.

Equation 41 implies that to maximise the margin is equivalent to minimise the norm $||\mathbf{w}||$.

$$g(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + b = d_i \Leftrightarrow r = \frac{g(\mathbf{x}_i)}{||\mathbf{w}||} = \frac{d_i}{||\mathbf{w}||} \tag{41}$$

Maximisation Problem 39 may now be formulated as Minimisation Problem 42.

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \mathbf{w}^T\mathbf{w} \\ & s.t.\ \mathbf{w}^T\mathbf{x_i} + b \geq 1 \qquad \forall i : d_i = +1 \\ & s.t.\ \mathbf{w}^T\mathbf{x_i} + b \leq -1 \qquad \forall i : d_i = -1 \end{aligned} \tag{42}$$

4.3.2 Non-linearly separable SVM

There are classification problems that fail to be linearly separable. In these cases, one way to solve the problem while still taking advantage of the amiable properties of hyperplanes, compared to a general shape, is to map the features from the input space $\mathbb{R}^n$ into a higher dimensional space $\mathbb{R}^m$, where $m > n$, by a function $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ where the points are linearly separable.[24]

Conceptually, this is illustrated in Figure 2, where a set of points that are not linearly separable in $\mathbb{R}$ are mapped with a non-linear map, $x \to (x, x^2)$, to $\mathbb{R}^2$ where they become linearly separable.

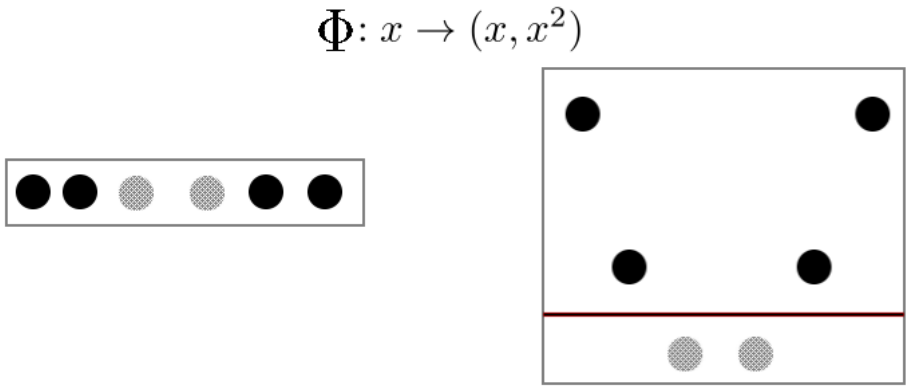

Figure 2: *The left figure shows a set of points in $\mathbb{R}$ which are not linearly separable. In the right figure the points are mapped to $\mathbb{R}^2$ by $\Phi : x \to (x, x^2)$, where they become linearly separable.*

Minimisation Problem 42 is then straightforwardly reformulated as Minimisation Problem 43. Remark that the weights, $\mathbf{w}$ will now be in the higher dimensional space.

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \mathbf{w}^T\mathbf{w} \\ & s.t.\ \mathbf{w}^T\Phi(\mathbf{x_i}) + b \geq +1 \qquad \forall i : d_i = +1 \\ & s.t.\ \mathbf{w}^T\Phi(\mathbf{x_i}) + b \leq -1 \qquad \forall i : d_i = -1 \end{aligned} \tag{43}$$

In practise, it quickly becomes computationally intense to map the features to a higher dimensional space so had it not been for the *Kernel trick* the SVM would hardly be feasible. The Kernel trick comes from noting that in Minimisation Problem 43, the features are included only as a dot product with the weights,

[24] Φ will be used throughout this thesis as the function that maps a feature into a higher dimensional space. When applied on a vector it should be interpreted as the mapping of the vector to the higher dimensional space. When applied on a matrix it should be interpreted as the map for all of its columns to the higher dimensional space.

and the weights only in dot products with the features and itself. This is the Kernel trick, from which the following observations may be made:

- Knowing the dot product between the weights and features will be sufficient, there is no need to actually calculate the weights and features explicitly.
- It is possible to let features and their corresponding weights be in infinite dimensional space given that the dot product is valid and possible to calculate.
- The map Φ does not have to be explicitly constructed; it is sufficient that the dot products between $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$, and $\Phi(\mathbf{x}_i)$ and the weights $\mathbf{w}$ are known for all i, j.

The *Kernel matrix* is defined as the dot product between the features in the higher dimensional space created by the map Φ, see 44.

$$\mathbf{K} = \Phi(\mathbf{x})^T \Phi(\mathbf{x}) \tag{44}$$

Each element in the Kernel matrix is a dot product of features in the higher dimensional space, and the dot product is referred to as the *Kernel function* and denoted as in 45. Remark that the Kernel function and thus Kernel matrix is symmetric.

$$k(\mathbf{x}_i, \mathbf{x}_i) = k_{ij} = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \tag{45}$$

In practise, it is usually more common to first pick a Kernel function, and do not care too much about what the map, Φ, looks like. However not all Kernels are valid to use if one wants to rely on the standard derivation of SVM that has been done here. For a Kernel to be valid, it has to be a dot product in some space.[25] This is equivalent to that the Kernel fulfils *Mercer's theorem*. Mercer's theorem is a theorem from functional analysis that in this context tells for a given function whether there is some space in which that function is a dot product.[26]

4.3.3 Soft margin formulation

In reality, a 100% prediction accuracy possibility on new data is seldom or never possible. There will always be some observations that will see an event against all odds - even a billionaire may default on his mortgage. However, it is certainly possible to achieve a 100% prediction accuracy on already known data with a complicated enough decision boundary. In the SVM case, by mapping to a space with sufficiently many dimensions, even a set of points that got their labels randomly picked out, so the features bear no relationship with the labels, will be separable.

[25] However, practitioners have nonetheless used SVMs with non-valid Kernels with success.
[26] Haykin et al., *Neural networks and learning machines*, p. 283.

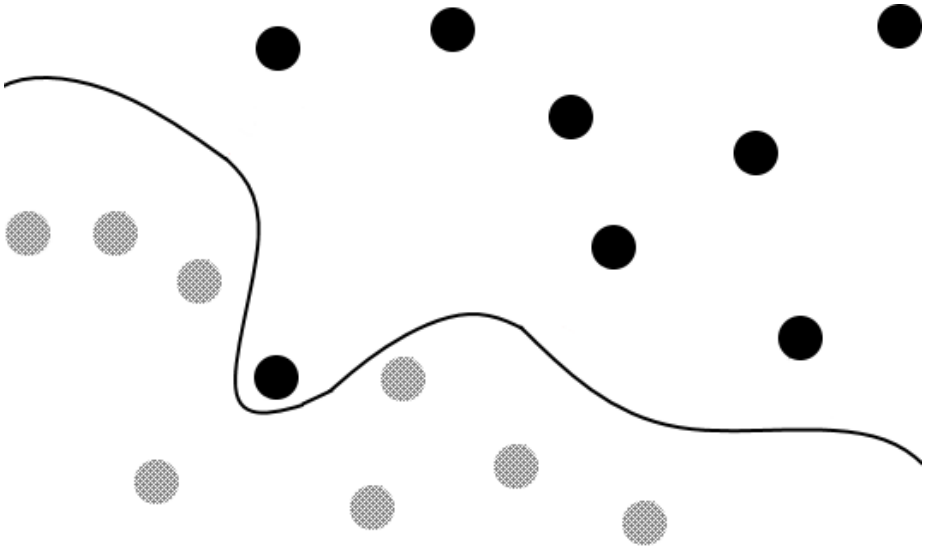

Figure 3: *A set of points with a decision boundary taken from higher dimensional space. The decision boundary has clearly been overfitted.*

That it is possible to achieve a 100% prediction accuracy on training data does not mean that it is a good idea. In Figure 3, an SVM has been trained in a higher dimensional space. One of the black points in Figure 3 is very close to the grey ones, but is still classified among the black ones. This means that another point that is placed on the same position is predicted to also be black. However, it seems more prudent and likely that the decision boundary in Figure 4 will generalise to new data.

The idea behind a soft margin is to allow for misclassified points, but give them a penalty in the objective function. The objective function in Minimisation Problem 43 is therefore changed to the one in Objective Function 46.

$$\min_{\mathbf{w},b} \quad \mathbf{w}^T\mathbf{w} + C\sum_i L(d_i, \mathbf{x}_i) \tag{46}$$

In 46, L is a loss function which imposes a penalty for misclassified points. C is a constant, or a *hyperparameter* that determines the trade off between minimising $\mathbf{w}^T\mathbf{w}$ and the loss.[27] Common choices for the loss function are $L(d_i, \mathbf{x}_i) = max(0, -sign(f_i)\cdot sign(d_i))$, $L(d_i, \mathbf{x}_i) = max(0, -f_i d_i)$ and $L(d_i, \mathbf{x}_i) = (f_i - d_i)^2$ where as before, f_i is the signed distance to the hyperplane and d_i the label. The last one of these loss functions is referred to as *least square penalty*. The remaining theory and the application of SVM in this theory will use least square penalty, since it brings a more straightforward solution than other types of loss

[27] A hyperparameter is a parameter set before training, in contrast with parameters or weights learned in the model training.

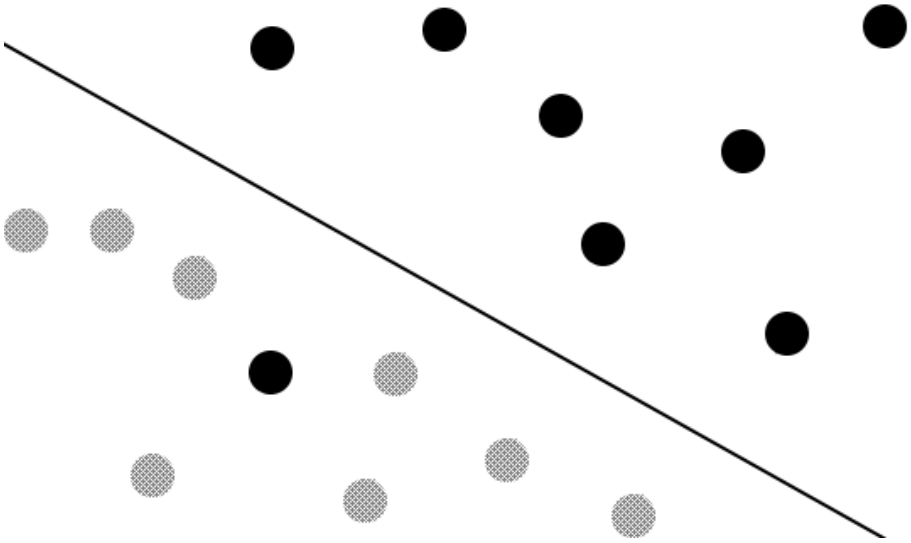

Figure 4: *The same set of points as in Figure 4. Even though one point is misclassified, this classification seems more likely to generalise to new data.*

functions, has good performance and will enable for an easy move to the SVR model which will be introduced afterwards.

4.3.4 Solution to optimisation problem

Since least square penalty will be used, the minimisation problem becomes the one in 47 where the vector $\mathbf{u}$ contains the errors.

$$\begin{aligned} \underset{\mathbf{w},b}{min} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}C \cdot \mathbf{u}^T\mathbf{u} \\ & s.t.\ \mathbf{w}^T\Phi(\mathbf{x_i}) + b + u_i = y_i \qquad \forall i \end{aligned} \tag{47}$$

Now that the general optimisation problem for least squares SVM has been stated, it will be solved. The optimisation problem to be solved is stated in Minimisation Problem 47. The problem is convex so it will be solved with standard convex optimisation routines. The Lagrangian for Minimisation Problem 47 is formulated in 48 where α_i are the Lagrangian multipliers, or *support vectors* as they are referred to in the SVM context.

$$\mathcal{L} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}C \cdot \mathbf{u}^T\mathbf{u} - \sum_i \alpha_i(\mathbf{w}^T\Phi(\mathbf{x_i}) + b + u_i - y_i) \tag{48}$$

Since the problem is convex, it can be solved by differentiating the Lagrangian from 48 w.r.t. $\mathbf{w}, b, \mathbf{u}$ and $\boldsymbol{\alpha}$ and setting the derivatives to zero, which yields

Equations 49.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \quad &\rightarrow \quad \mathbf{w} = \boldsymbol{\alpha}^T \Phi(\mathbf{x}) \\
\frac{\partial \mathcal{L}}{\partial b} = 0 \quad &\rightarrow \quad \mathbf{1}^T \boldsymbol{\alpha} = 0 \\
\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = 0 \quad &\rightarrow \quad \boldsymbol{\alpha} = C\mathbf{u} \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = 0 \quad &\rightarrow \quad \mathbf{y} = \Phi(\mathbf{x})^T \mathbf{w} + \mathbf{1} \cdot b + \mathbf{u}
\end{aligned} \tag{49}$$

Eliminating $\mathbf{w}$ and $\mathbf{u}$ with the first and third equation in 49 lessens the number of equations to the two in Equations 50. As before, $\mathbf{K}$ is the Kernel matrix, defined as $\Phi(\mathbf{x})^T \Phi(\mathbf{x})$.

$$\begin{aligned}
&\mathbf{1}^T \boldsymbol{\alpha} = 0 \\
&\mathbf{y} = \Phi(\mathbf{x})(\alpha^T \Phi(\mathbf{x})) + \mathbf{1} \cdot b + \boldsymbol{\alpha} \cdot C^{-1} = (\mathbf{K} + \mathbf{I} \cdot C^{-1})\boldsymbol{\alpha} + \mathbf{1} \cdot b
\end{aligned} \tag{50}$$

The equations in Equation 50 can be summarised in Equation System 51.

$$\begin{pmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \mathbf{K} + C^{-1} \cdot \mathbf{I} \end{pmatrix} \begin{pmatrix} b \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix} \tag{51}$$

From the second row in Equation 51 it can be seen that $\boldsymbol{\alpha} = (\mathbf{K} + C^{-1} \cdot \mathbf{I})^{-1}\mathbf{y} - \mathbf{1} \cdot b$ and combing this with the first row yields: $\mathbf{1}^T \cdot (\mathbf{K} + C^{-1} \cdot \mathbf{I})^{-1}(\mathbf{y} - \mathbf{1} \cdot b) = 0$ which gives Equation 52.

$$b\mathbf{1}^T \cdot (\mathbf{K} + C^{-1} \cdot \mathbf{I})^{-1}\mathbf{1} = \mathbf{1}^T \cdot (\mathbf{K} + C^{-1} \cdot \mathbf{I})^{-1}\mathbf{y} \tag{52}$$

The value of b is thus obtained by 53:

$$b = \frac{\mathbf{1}^T \cdot (\mathbf{K} + C^{-1} \cdot \mathbf{I})^{-1} \cdot \mathbf{y}}{\mathbf{1}^T \cdot (\mathbf{K} + C^{-1} \cdot \mathbf{I})^{-1} \cdot \mathbf{1}} \tag{53}$$

When combining 53 with 51 the value of $\boldsymbol{\alpha}$ is therefore the one in 54.

$$\boldsymbol{\alpha} = (\mathbf{K} + C^{-1} \cdot \mathbf{I})^{-1} \cdot (\mathbf{y} - b \cdot \mathbf{1}) \tag{54}$$

4.3.5 Prediction

Since $\mathbf{w}$ has been solved for in terms of $\boldsymbol{\alpha}$ and $\Phi(\mathbf{x})$, there is no need for calculating $\mathbf{w}$ explicitly. Let i be a new observation. The definition of how to calculate f_i is Equation 55.

$$f_i = \mathbf{w}^T \Phi(\mathbf{x}_i) + b \tag{55}$$

Since $\mathbf{w} = \boldsymbol{\alpha}^T \Phi(\mathbf{x})$, Equation 55 can be calculated as in Equation 56, where j represents the points in the training data.

$$\mathbf{w}^T \Phi(\mathbf{x}_i) + b = \boldsymbol{\alpha}^T \Phi(\mathbf{x}) \Phi(\mathbf{x}_i) + b = \sum_j \alpha_j k_{ij} + b \tag{56}$$

As defined before, a new observation is classified into 1 if $f_i \geq 0$, and -1 if $f_i < 0$.

4.3.6 Radial Basis Kernel

One of the most common choices of kernels is the *Gaussian Radial Basis Kernel* or just *Radial Basis Kernel*. A radial basis kernel in general is a kernel which only depends on the features through their distance, and the Gaussian function is the by far most common. The proof that the kernel is valid will not be stated here but it is well established that it is valid.[28] The kernel function is stated in Equation 57.

$$k_{ij} = e^{-\frac{1}{2\sigma^2} ||\mathbf{x}_i - \mathbf{x}_j||^2} \tag{57}$$

The interpretation of the kernel is as a similarity measure. σ^2 is a hyperparameter, similar to C, and has to be fixed before training the model. It will later be discussed how to choose σ^2. The maximum value is 1 when $\mathbf{x}_i = \mathbf{x}_j$ and the limit is 0 when $||\mathbf{x}_i - \mathbf{x}_j|| \to \infty$. The kernel corresponds to a dot product in an infinite dimensional space, meaning that the function Φ maps all the features to an infinite dimensional space where the dot product of the features is equal to the kernel function in Equation 57.

4.3.7 Tuning of hyperparameters

The value of C should be chosen to be the one that gives the best predictions. When evaluating the prediction, this cannot be done on the data used to calibrate the model, referred to as the *training set*, because it will promote overfitting. If the number of different observations in the training set are p, they can be perfectly separated in $p + 1$ dimensions, and overfitting may occur well before that. The quality of the model with different values of C will therefore be evaluated on *validation data*, which is a completely new set of data. In the validation data, an overfitted model will perform worse than a well fitted one.[29] The results of the fit on the training and validation data may look like Figure 5.

However since C is chosen with basis on the validation data, it gets biased towards fitting that data. The expectation of the relationship between fit and

[28] Haykin et al., *Neural networks and learning machines*, p. 284.
[29] Ibid., p. 38-39.

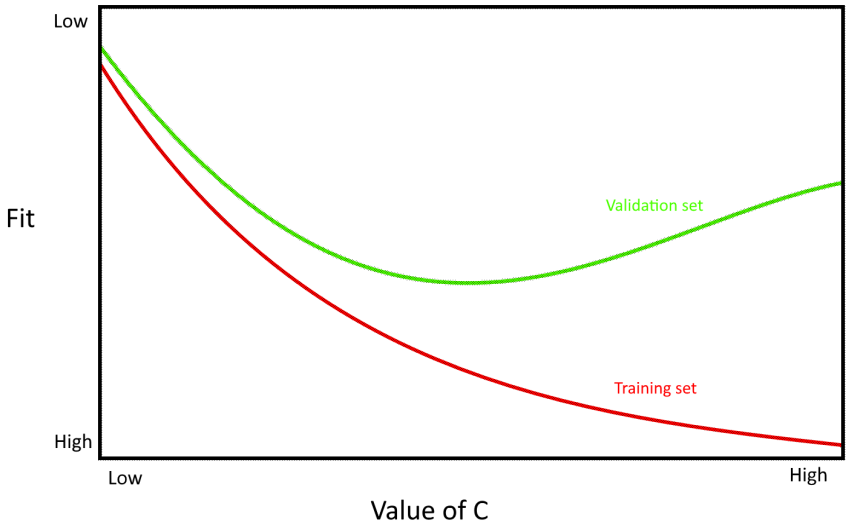

Figure 5: *The expected relationship between fit and C. A high C means a low tolerance for misclassified points in the training data.*

value of C may be smooth as in Figure 5, but becomes noisy in practise, more like Figure 6. As can be seen in the figure, the value of C that is chosen, happens to be one in a valley, which is probable made up by noise. When C has been chosen, the evaluation of prediction accuracy should therefore be done on new data, referred to as *test data*.

Like C, the hyperparameter from the radial basis kernel, σ^2, will be chosen by cross-validation on the validation data. The choices of C and σ^2 may not be independent, meaning that the value for C that gives the best fit is dependent on which value of σ^2 that has been chosen, and vice versa. A standard way of finding the optimal hyperparameters is to do a *grid search*. This means to train the model on the training data with different values of the hyperparameters, and afterwards calculate the fit on the validation data.

Usual starting values for the hyperparameters in the grid are different powers of 2, both negative and positive, although more towards more positive ones.[30] This thesis will start with a grid search of σ^2 and C in the range $(C, \sigma^2) \in \{2^i, 2^j\}_{(i,j)=(-7,-7)}^{(12,12)}$. A qualitative check will then be done to see whether it seems likely the optimal values are in that range. If so, the values with the best prediction will be chosen. However, if it seems like the range is more towards the boundary of the grid, a new grid search will be done in near that boundary where the procedure will be repeated.

[30] Hsu et al., 'A practical guide to support vector classification'

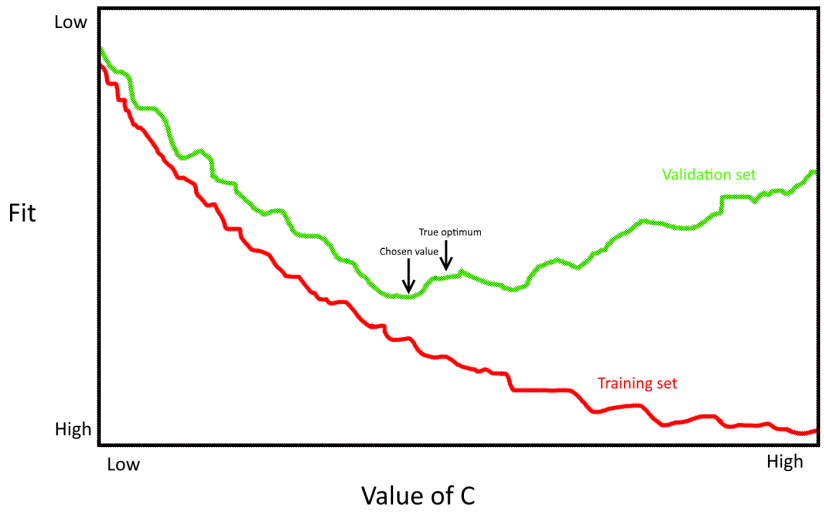

Figure 6: *A possible empirical relationship between fit and C. The choice of C will get biased to the validation data, so the final performances should be evaluated on another data.*

4.3.8 Platt's probabilistic scaling

An SVM works as a binary classifier and does not provide probabilities in the classification. The standard use of SVM is for classifying points into two categories, and as such it has been mostly used for problems where an observation shall be placed in one of two buckets, which determines whether an action will be taken. An example of this could be whether to approve a loan application or not.

However in the context of IFRS 9, the interest lays in the expected credit loss which implies that probabilities for the observations to be classified into the buckets are needed. One way to do this is called *Platt's probabilistic scaling* which was proposed by John C. Platt in 1997.[31] The model makes use of the information from the support vector machine. It assumes that the log-odds of being classified into a category has a linear relationship with the distance from the decision boundary, i.e. it is a logistic regression with the labels as response and the distance to the decision boundary from the SVM as the explanatory variable. The main advantage of the model is that it is a simple but yet powerful way of making use of the SVM predictions to produce probabilities.

The estimated probability function for an observation i will be Equation 58.

[31] Platt, 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods'.

$$1 - P(d_i = -1) = P(d_i = 1) = \frac{1}{1 + e^{-(A+Bf_i)}} \tag{58}$$

where A and B are calibrated by an ordinary logistic regression model. See Figure 7 for an example of how new points are assigned with a probability.

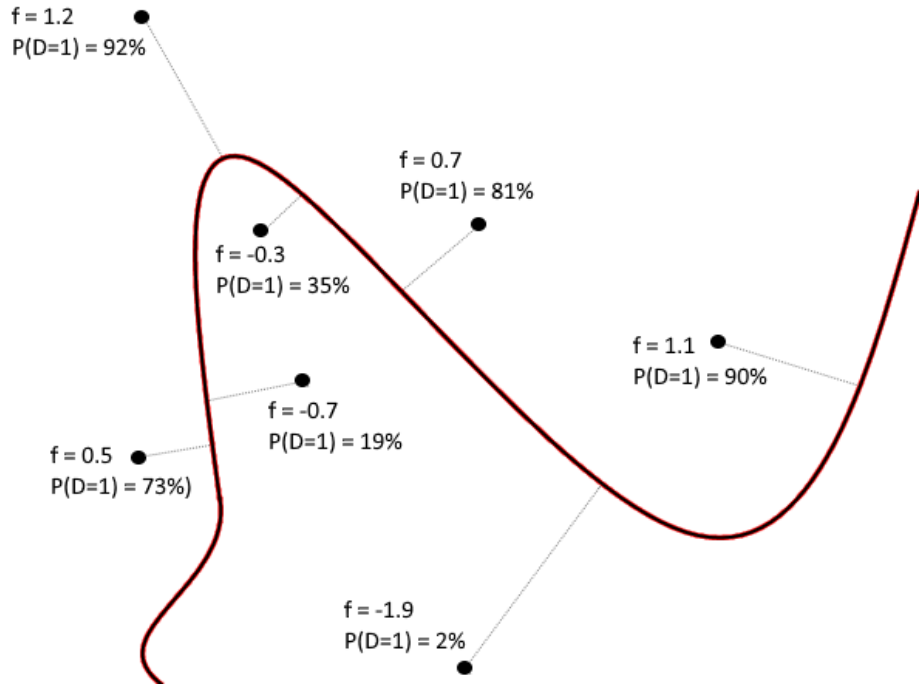

Figure 7: *The figure shows an example of predicting probabilities by converting f_i to probabilities for new observations after a logistic regression has been trained on f_i from training data*

A straightforward generalisations is to divide f into $f^+ = max(f, 0)$ and $f^- = min(f, 0)$ and use both as explanatory variables in the logistic regression, which yields Model 59. This is more general than Platt's probabilistic scaling since it allows for a different probability relationship on the different sides of the margin, which could improve performance.

$$1 - P(d_i = -1) = P(d_i = 1) = \frac{1}{1 + e^{-(A+Bf_i^+ + Ef_i^-)}} \tag{59}$$

4.4 Support Vector Regression

Support Vector Regression (SVR) is a model that is distinct, but similiar to SVM. The main difference is that it it used for problems with a continuous

dependent variable, like an ordinary linear regression. However, it is very similar to SVM in the way that it has a setup using convex optimisation, and uses a mapping to higher dimensional space to account for complicated non-linear relationship.

Intuitively, SVR maps the features to a higher dimensional space. In the higher dimensional space, it tries to find the hyperplane that best fits the dependent variables, similar to an ordinary linear regression with a specified loss function. In this thesis, the loss function will always be least square, meaning that it is very similar to a linear regression, albeit in another space.

The main difference from a linear regression, aside from the mapping to higher dimensions, is that the function that will be minimised is dependent on the least square loss, but also on $\mathbf{w}^T\mathbf{w}$. This is increasing in higher absolute values of the elements in the weights. The reason for this is that when mapping the regression hyperplane back to input space, it will correspond to a more complicated boundary if the norm of the weights is larger. It is thus a penalty for a complicated surface.

The optimisation problem for least square SVR is stated in Minimisation Problem 60. As before, $\mathbf{u}$ is the vector of errors and C is the constant to determine the trade off between a complicated regression hyperplane and few misclassifications in the training data. C will be chosen by cross-validation with the same method as for SVM, which was described in section 4.3.7.

$$\begin{aligned} \underset{\mathbf{w},b}{min} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}C \cdot \mathbf{u}^T\mathbf{u} \\ & s.t.\ \mathbf{w}^T\Phi(\mathbf{x_i}) + b + u_i = y_i \qquad \forall i \end{aligned} \tag{60}$$

Minimisation Problem 60 is exactly the same as Minimisation Problem 47 in the least square SVM model. This means that the derivations of the solutions are the same and that the estimation and prediction are the same as for the least square SVM model stated in the previous section.

5 Methodology

This section will first describe how the data will be used in the different approaches, then proceed to a description and rationale for the models, and sum up with the methods to be used for the evaluation of the approaches. Calibration refers to estimating the models and their parameters based on the training data, while prediction refers to calculating the expected LGD on the test data, which is then to be compared with the real values.

As for a reminder, all approaches will partion LGD into a probability of loss and a conditional severity, see Equation 61, which was derived in Equation 3 in the introduction. Here $LGD \neq 0$ has been changed to $LGD > 0$ since no negative LGDs were used, as said in section 3.

$$E(LGD) = \underbrace{P(LGD > 0)}_{\text{Probability of loss}} \cdot \underbrace{E(LGD|LGD > 0)}_{\text{Conditional severity}} \tag{61}$$

5.1 Data

Totally, eight years of data will be used in this thesis, consisting of LGDs with corresponding explanatory variables from 2008 to 2015. The last two, i.e. 2014-2015 will be used as a test set on which the approach will be evaluated. The rationale for the choice of this period for the test data is that LGDs are generally predicted for the current exposures, but can only be calibrated on already defaulted ones. This implies that calibration should be done on data in another time range than the prediction when assessing the fit.

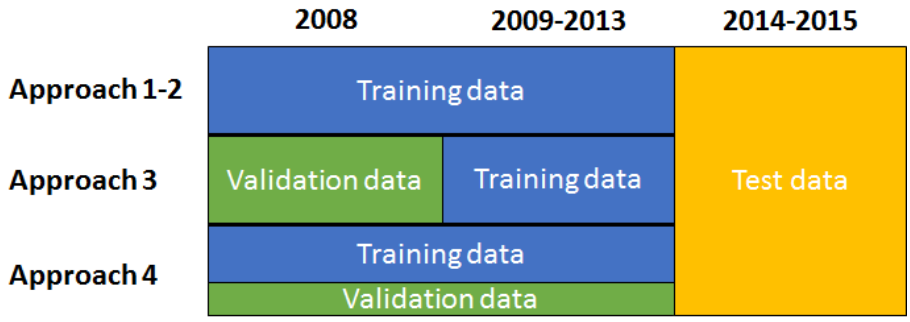

Figure 8: *The eight years of data that will be used are divided into training, validation and test sets for the different approaches.*

The rest of the data will be used to estimate the parameters in the approaches. For Approach 1 and 2 this means that all the remaining data, that is 2008-2013, will be training data. In Approach 3, 2008 will be reserved for validation in order to tune the hyperparameters. 2008 is picked to have a year that is outside the

time range of the training data, but that still lets the training data be adjacent to the time period of the test data. In Approach 4, the validation data will be 25% randomly picked observations from 2008-2013, while the rest of the data from 2008-2013 will be training data. Since time is not included as a feature in Approach 4, there is no need for setting aside a whole year for validation purposes. See Figure 8 for an overview of the division of the data.

5.2 Approach 1: Standard regressions

This approach consists of a logistic regression for predicting the probability that a loss will be greater than zero, and a linear regression predicting the expected conditional loss severity. The linear and logistic regressions will be calibrated on the training data with standard techniques, which yields the estimated coefficients $\hat{\boldsymbol{\beta}}_S$ and $\hat{\boldsymbol{\beta}}_P$. The prediction is then calculated as in Equation 62.

The prediction is calculated as in Equation 62 where $\mathbf{x}$ is the explanatory variables.

$$\begin{aligned} E_{\hat{\boldsymbol{\beta}}_S,\hat{\boldsymbol{\beta}}_P}(LGD) &= E_{\hat{\boldsymbol{\beta}}_S}(LGD|LGD>0)\cdot P_{\hat{\boldsymbol{\beta}}_P}(LGD>0) \\ &= \underbrace{\hat{\boldsymbol{\beta}}_S^T\cdot\mathbf{x}}_{\text{Conditional severity}} \cdot \underbrace{\frac{1}{1+exp(-(\hat{\boldsymbol{\beta}}_P^T\cdot\mathbf{x}))}}_{\text{Loss Probability}} \end{aligned} \tag{62}$$

The approach is simple and standard which makes it easy to implement and explain, and is widely used in the market.[32] A drawback with the approach is that if there are non-linear effects or variable interactions then the model could lose explanatory power compared to more advanced approaches. The approach is not of interest in itself but is included as a way of benchmarking the other approaches.

All variables in Table 1 except time will be used for both the linear and logistic regressions. The reason that time is not included is that it would not be a prudent approach when predicting LGD on data where the time periods are in a different range than the time periods in the training data. This is because a local trend in increasing or decreasing LGD with respect to time with this model will become a global trend in the prediction, making the LGD go to improper values in the long term.

APPROACH 1

A summary of calibration and prediction in Approach 1

1. **Calibrate the model**

[32] Zhang and Thomas, 'Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD'.

(a) Calibrate standard linear and logistic models for obtaining the parameter estimates

(b) Save the parameters $\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\beta}}_P$

2. **Predict LGDs on new data**

(a) Calculate probability of loss according to the estimated parameters

(b) Calculate expected conditional severity according to the estimated parameters

(c) Calculate predicted LGD by multiplying probability of loss with expected conditional severity

5.3 Approach 2: Regressions with time varying intercept

Approach 2 is similar to Approach 1 but contains the time varying intercept described in the Theoretical Framework section 4.2. The rationale for the model is that the LGD figures change over time, and some but not all of this may be explained by macro economic variables. The approach is flexible in the sense that it will take into account all idiosyncratic information but quickly adopt itself to a new level if for example business practise or new regulations change the average LGD values by a parallel shift.

The approach assumes that the loss probability follows a logistic regression model and that the conditional severity follows a linear regression model, conditioned on the time varying intercept for each model respectively. The intercept is assumed to move between time periods according to a normal distribution with a drift that depends on macro economic variables and a constant variance. This means that the time varying intercept, α_t, changes over time periods according to Equation 63.

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1} + \boldsymbol{\theta}^T \mathbf{z}_t, \sigma_\alpha^2) \tag{63}$$

When predicting values of LGDs with the time varying intercept approach, the most correct way would be to first calculate the conditional prediction given all possible values of α for that time period, and then integrate over the probability distribution of α, which has been derived in the theoretical framework. However, this would complicate the prediction a lot, and the gain is assessed to be small. Instead the prediction will condition on $\boldsymbol{\alpha} = E(\boldsymbol{\alpha})$.

Let N as before be the latest time period in the training data, and $\mathbf{z}$ the macro economic variables. Now to calculate the expected value of α_K where $K > N$. From Equation 63, it can be seen that the distribution of α_K conditioned on

α_N is independent of α_J where $N > J$. The expected value of α_N has already been derived in the calibration and from this the expected value of α_K can be calculated by Equation 63 where the notation $E(\alpha_t) = \hat{\alpha}_t$ is used.

$$\begin{aligned} E(\alpha_K|\alpha_N) &= \boldsymbol{\theta}^T \mathbf{z}_K + E(\alpha_{K-1}|\alpha_N) \\ &= \sum_{t=N+1}^{K} \boldsymbol{\theta}^T \mathbf{z}_t + E(\alpha_N|\alpha_N) \\ &= \sum_{t=N+1}^{K} \boldsymbol{\theta}^T \mathbf{z}_t + \hat{\alpha}_N \end{aligned} \tag{64}$$

A LGD for period K is thus predicted as in Equation 65, where $\hat{\boldsymbol{\alpha}}_{S,K}$ and $\hat{\boldsymbol{\alpha}}_{P,K}$ denotes the estimated intercept for period K for the linear and logistic regression respectively.

$$\begin{aligned} E_{\hat{\boldsymbol{\beta}}_S,\hat{\boldsymbol{\theta}}_S,\hat{\boldsymbol{\alpha}}_{S,K},\hat{\boldsymbol{\beta}}_P,\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\alpha}}_{P,K}}(LGD) &= E_{\hat{\boldsymbol{\beta}}_S,\hat{\boldsymbol{\theta}}_S,\hat{\boldsymbol{\alpha}}_{S,K}}(LGD|LGD>0) \cdot P_{\hat{\boldsymbol{\beta}}_P,\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\alpha}}_{P,K}}(LGD>0) \\ &= \underbrace{\left(\hat{\boldsymbol{\beta}}_S^T \mathbf{x} + \hat{\boldsymbol{\alpha}}_{S,K}\right)}_{\text{Conditional severity}} \cdot \underbrace{\frac{1}{1+exp(-(\hat{\boldsymbol{\beta}}_P^T \mathbf{x} + \hat{\boldsymbol{\alpha}}_{P,K}))}}_{\text{Loss Probability}} \end{aligned} \tag{65}$$

APPROACH 2

A summary of calibration and prediction in Approach 2

1. **Calibrate the model**
   (a) Calibrate standard linear and logistic models to be used for starting values of the parameters, i.e. the parameters from Approach 1
   (b) Repeat Expectation-Maximisation algorithm until convergence (defined as the first step all parameters have changed less than 0.1%) for both the linear and logistic model
      i. Calculate the distribution of the time varying intercept variables given current parameter values
      ii. Calculate the expected value of the log likelihood given the current estimate distribution of the time varying intercept
      iii. Find the new parameters by maximising the expected likelihood from step (b)
   (c) Save the parameters $\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\beta}}_P, \hat{\boldsymbol{\theta}}_P$
   (d) Save the expected value of α_T, i.e. the latest intercept in both the linear and logistic regression

2. **Predict LGDs on new data in time period K**

   (a) Calculate the time varying intercept for probability of loss by Equation 64

   (b) Calculate probability of loss according to the estimated parameters and the time varying intercept

   (c) Calculate the time varying intercept for expected conditional severity by Equation 64

   (d) Calculate expected conditional severity according to the estimated parameters and the time varying intercept

   (e) Calculate predicted LGD by multiplying probability of loss with expected conditional severity

5.4 Approach 3: Support Vector Machine & Regression with time

Approach 3 is similar to Approach 1 and 2 in the sense that LGD is divided into probability of loss and conditional severity but is otherwise quite different. Instead of employing linear and logistic regression, it uses SVM, with a version of Platt's probability scaling to model loss probability, and SVR to model loss severity.

This approach includes all the explanatory variables in Table 1 as features, including time. Time is a very special variable since the test data, and data in general that is of interest for prediction, lay in the future meaning that it will contain values on the variable time in a range disjoint from the training data. Since the radial basis kernel that is used for both the SVM and SVR enacts as a similarity measure, it is uncertain whether it is at all possible to get reasonable predictions on data where the time is in another range than the training data.

When estimating the model, the first step is to determine the hyperparameters C and σ^2 with cross-validation by a grid search on the validation data. The search will start in the range $\{2^k\}_{k=-7}^{12}$ for both C and σ^2. The models will be calibrated with all values of the hyperparameters in this grid. Then the models will predict on the validation data. For SVR, the RMSE on the validation data will be recorded for all cells in the grid. Here a qualitative selection will be made.

For a new observation, the conditional severity is predicted by the SVR by using the formula in Equation 66 derived in Equation 56 in the framework. The support vectors α are estimated on the training data with the chosen hyperpa-

rameters. Denote this prediction for conditional severity f_S.

$$f = \sum_j \alpha_j k_{ij} + b \tag{66}$$

The prediction for the probability of loss is calculated by first calculating the prediction as in Equation 66. Denote this prediction f_P. From this, the probability of a loss is calculated by Platt's scaling, with the generalisation that used f^+ and f^- as regressors rather than f. The probability of loss will thus be estimated by Equation 67 where $\hat{A}$, $\hat{B}$ and $\hat{E}$ are estimated coefficients in the logistic regression.

$$P(d = 1) = \frac{1}{1 + e^{-(\hat{A} + \hat{B} f_P^+ + \hat{E} f_P^-)}} \tag{67}$$

The prediction for LGD is then calculated as in Equation 68 where $\boldsymbol{\alpha}$ are the support vectors.

$$E_{\boldsymbol{\alpha}_S, \boldsymbol{\alpha}_P, \hat{A}, \hat{B}, \hat{E}, C, \sigma^2}(LGD) = \underbrace{f_S}_{\text{Conditional severity}} \cdot \underbrace{\frac{1}{1 + exp(-(\hat{A} + \hat{B} f_P^+ + \hat{E} f_P^-))}}_{\text{Loss Probability}} \tag{68}$$

APPROACH 3 & 4

A summary of calibration and prediction in Approach 3 & 4

1. **Calibrate the SVM and SVR models**
   (a) Choose values of the hyperparameters C and σ^2 for both the SVM and SVR model according to the methodology in the theoretical framework section 4.3.7
   (b) Estimate the support vectors $\boldsymbol{\alpha}_S$ on the training data according to the analytical solution for the least square SVM model
   (c) Estimate the support vector $\boldsymbol{\alpha}_P$ on the training data according to the analytical solution for the least square SVR model
2. *Calibrating logistic regression on SVM output for obtaining probabilities*
   (a) Calculate the predicted value of f on the validation set by Equation 66
   (b) Estimate parameters for obtaining probabilities by setting up a linear regression with loss event as dependent variable and f^+ and f^- as explanatory variables

(c) Save the parameters from this logistic regression as $\hat{A}, \hat{B}$ and $\hat{E}$

3. **Predict LGDs on new data**

   (a) Calculate the predicted value for loss severity, f_S on new data using Equation 66

   (b) Calculate the predicted value for probability of loss, f_S on new data using Equation 66

   (c) Calculate probability of loss by applying the estimates from the logistic regression, $\hat{A}, \hat{B}$ and $\hat{E}$, to f_P^- and f_S^+

   (d) Calculate the predicted LGD by multiplying probability of loss with expected conditional severity

5.5 Approach 4: Support Vector Machine & Regression without time

This approach is completely analogous to Approach 3 except for that time is not included as a feature. It uses exactly the same method for calibration and prediction. The only differences except that time as a variable is excluded is that the validation data instead of being one year, now is 25% of the data taken completely randomly in the training data. The reason for separating the two approaches is that time is a very special kind of variable so having one approach without time will make it easy to compare and see how the models will adjust.

5.6 Model evaluation

This section will describe the statistical tests that will be carried out for testing the model assumptions in Approach 1 & 2 and how the model prediction accuracy for all approaches will be evaluated.

5.6.1 Test for heteroscedasticity

The Breusch-Pagan test is used for detecting heteroscedasticity in linear regression models.[33] Let $\mathbf{y}$ be the response vector and $\mathbf{X}$ the design matrix. Let $\hat{\boldsymbol{\epsilon}}$ be the estimated residuals of a simple linear regression with parameters estimated by least squares. The residuals $\hat{\boldsymbol{\epsilon}}^2$ are fitted into a new regression model as

[33] Breusch and Pagan, 'A simple test for Heteroscedasticity and Random Coefficient Variation'.

response variable with $\mathbf{x}$ as explanatory variables, see Equation 69.

$$\hat{\epsilon}_i^2 = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i + \eta_i \tag{69}$$

The parameters in Equation 69 are estimated by ordinary least square fitting. The null hypothesis of homoscedasticity is tested against the alternative hypothesis of heterscedasticity. The coefficient of determination times the sample size is under the null hypothesis is approximately chi-square distributed with n degree of freedom, see Equation 70 where n denotes the number of explanatory variables, which gives a p-value for the test.

$$nR^2 \sim \chi_p^2 \tag{70}$$

5.6.2 Test for autocorrelation

Define $\bar{r}_t$ as the average LGD minus the average predicted LGD for the data period t. Autocorrelation will be searched for by inspecting the correlation between $\bar{\mathbf{r}}_{1:(N-1)}$ and $\bar{\mathbf{r}}_{2:N}$. The correlation will be weighted by number of observations averaged over the subsequent data periods.

5.6.3 Model prediction accuracy

There will be two evaluations of model prediction accuracy. The first one will be to calculate *Root Mean Square Error* (RMSE) for each customer in the test data. RMSE is a measure of fit, and defined as the root of the average of the squared residuals. This will be done for the models for conditional severity and loss probability as well as for the total predicted LGD. The purpose is to evaluate how much of the randomness that the models cannot explain.

The second evaluation of model prediction accuracy will be to check whether the average LGD for the two years in the test data matches the average predicted LGD for these years. From an IFRS 9 perspective, this is very important, since the purpose is to be correct on average. Individual mispredictions are not that much of a problem as long as they average each other out.

6 Results

In this section the results for the approaches will be presented. It will start with evaluating the parametric assumptions behind Approach 1 and 2, which will be followed by the tuning and choice of hyperparameters for Approach 3 and 4. Finally, the fit will be presented together with figures.

6.1 Parametric evaluation

As described in the Methodology section, the autocorrelation for the LGD in Approach 1 and 2 will be evaluated. There will also be a test for heteroscedasticity. Lastly there will be Q-Q-plots and an evaluation of the normality assumptions for the time varying intercepts.

6.1.1 Autocorrelation & Heteroscedasticity

The residuals are defined as the average of the observed LGDs minus the predicted ones (combining loss probability with conditional severity). The correlation between the average of the residuals in adjacent time periods was 16% for Approach 1 and -5% for Approach 2. This means that there is autocorrelation in the residuals from Approach 1, but not very much. That the autocorrelation was relatively small suggests that Approach 2 may not be such a big improvement compared to Approach 1. Though, it does show that most of the autocorrelation is removed with Approach 2, suggesting that the approach is sound from this perspective.

The result from the Breusch-Pagan test showed that there was heteroscedasticity in both the linear regressions in Approach 1 and 2. The p-value was less than 0.05 for Approach 1 and less than 0.01 for Approach 2. The Breusch-Pagan test also shows which variables had the most significant relationship with the squared residuals, which was the SME dummy variable in Table 1. This indicates whether a customer is a larger or smaller corporate. Since having this dummy variable caused heteroscedasticity, it suggest that larger corporates behave differently from smaller ones in a way which cannot be completely explained by a dummy variable. Even though there are some problems with heteroscedasticity, the linear regressions will still be unbiased and work relatively well, the lack of homoscedasticity is not therefore seen to be a big issue.

6.1.2 Normality of Alpha

The residuals, i.e. the random movement, in α is contained in the serie $\{\alpha_t - \alpha_{t-1} - \hat{\boldsymbol{\theta}}^T \mathbf{z}_t\}_t$. This serie is showed as a Normal Q-Q plot in Figure 9. As can be seen, the fitted residuals indicate a normal distribution, except for two

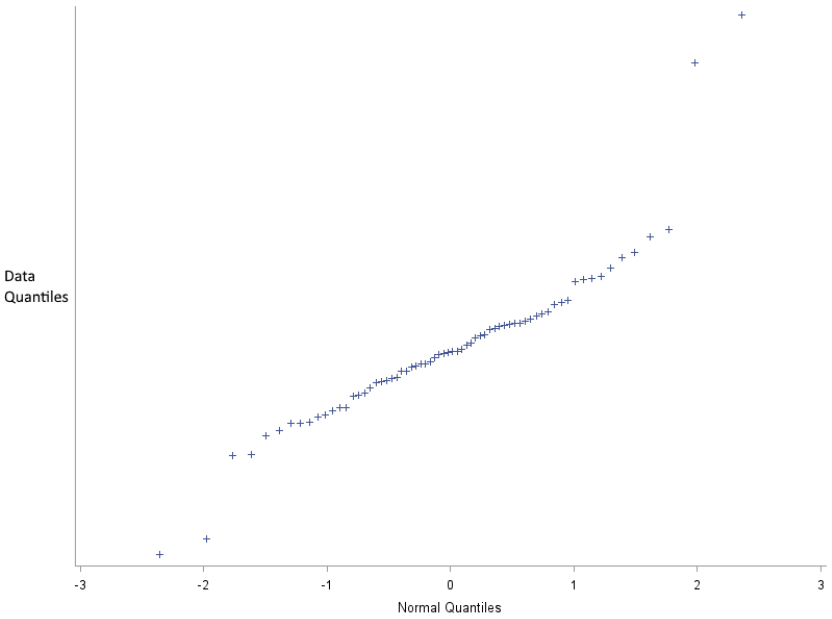

Figure 9: *Normal Q-Q plot for the residuals of the most likely α-values for the linear regression used for estimating loss severity.*

outliers on each end. There seems to be some tendency for extreme observations, however not enough to support changing the assumption of normal random movement to another distribution.

In Figure 10, the corresponding Q-Q plot for the random movement in the time varying intercept for the logistic regression is shown. While the plot for the linear regression showed normality except for the extremes, this plot shows a pattern much harder to interpret. At the upper quantiles, the end still looks normal. However the lower quantiles show a clear pattern of underdispersion, i.e. that the extremes are less extreme than what could be expected. A reason for this could be the logit-link in the logistic regression.

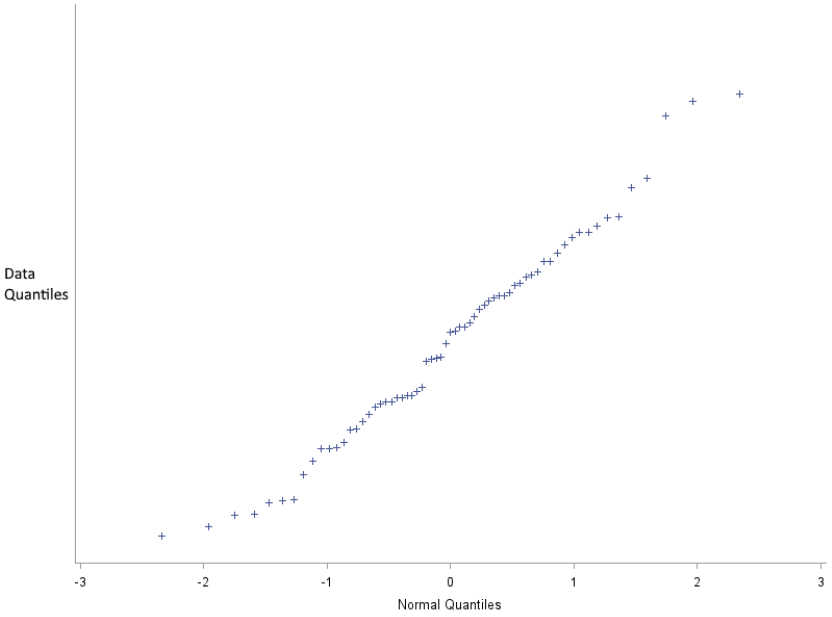

Figure 10: *Normal Q-Q plot for the residuals of the most likely α-values for the logistic regression used for estimating loss probability.*

6.2 Tuning hyperparameters

As described in 4.3.7, the hyperparameters C and σ^2 in Approach 3 and 4 will be tuned on the validation data. In Table 2 the hyperparameters chosen after tuning are presented for both approaches and their respective models. As can be seen, there is quite some variation, but the models with a large C tend to have a large σ^2 as well.

Table 2: *Chosen values of hyperparameters C and σ^2 for the SVM and SVR models in Approach 3 and 4*

Approach	Model	C	σ^2
3	SVM	2^{14}	2^{14}
3	SVR	2^3	2^5
4	SVM	2^{11}	2^9
4	SVR	2^{10}	2^7

In the rest of the section, the decision process of how the hyperparameters in Table 2 were chosen will be described.

6.2.1 Tuning of SVM in Approach 3

The SVM-model in Approach 3 has been trained 400 times for Figure 11, each time with different values of the hyperparameters. The cells in the figure represents how large the log-likelihood on the validation data was for that choice of hyperparameters. A green cell represents a high likelihood, relatively to the other, while a red cell represents a lower likelihood. Since the SVM model uses a lot of data, the runtime was very long on the full set. In order to decrease the computational time, the grid search for the hyperparameters was done on half of the training and validation data.

As can be seen, the region with the highest log-likelihood, and thus best fit, is the one to the bottom-right. Since the best fit seems to be near the boundary, the model was recalibrated with parameters in this region, and further to the bottom down, still with half of the data.[34] The region chosen for recalibration was $(C, \sigma^2) \in \{2^i, 2^j\}_{(i,j)=(5,6)}^{(19,16)}$, and the results are shown in Figure 12.

In Figure 12, the fit for the SVM model in Approach 3, recalibrated with new values of the hyperparameters are shown. The region where σ^2 was 12 and C was in the region 12-14 had the best fit. It was thus decided the calibrate the model on the whole data for the region $(C, \sigma^2) \in \{2^i, 2^j\}_{(i,j)=(12,12)}^{(15,15)}$, which is marked and dotted on the figure.

In Figure 13, the SVM model in Approach 3 has been calibrated on the whole training data, and the likelihood has been cross-evaluated on the whole cali-

[34] Here recalibration means that the model is calibrated again with the same data but other parameters.

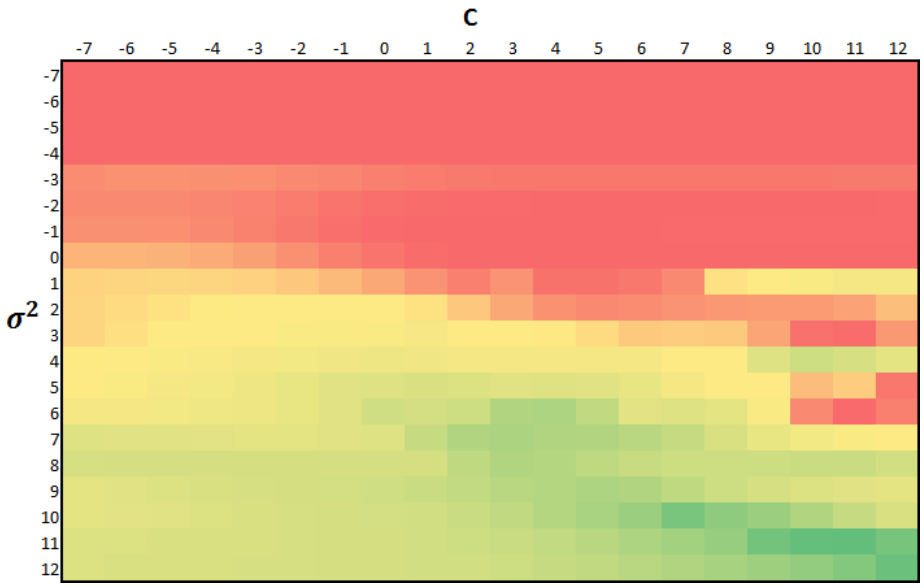

Figure 11: *Tuning of the hyperparameters in the SVM model in Approach 3 while using 50% of the data. Green represents a higher likelihood on the validation data while red represents a higher one. The scale is log_2 for both C and σ^2.*

bration data in the region with best fit when calibrating on half of the data. The hyperparameters which yielded the best fit were $(C, \sigma^2) = (14, 14)$, which therefore are chosen for the model.

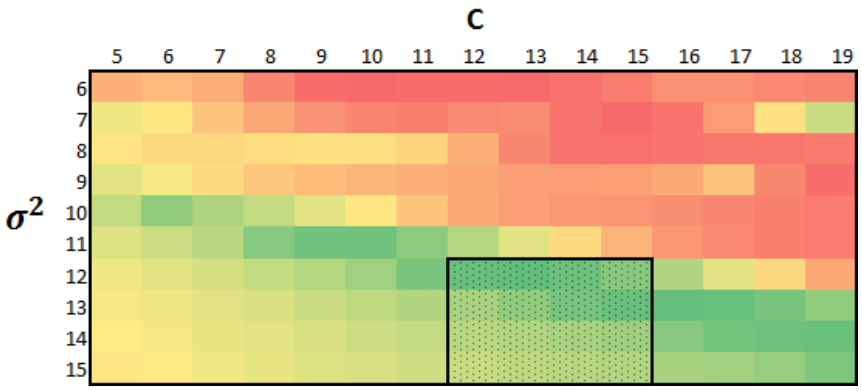

Figure 12: *Tuning of the hyperparameters in the SVM model in Approach 3 while using 50% of the data. Green represents a higher likelihood on the validation data while red represents a higher one. The scale is log_2 for both C and σ^2.*

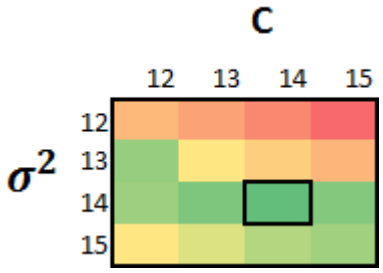

Figure 13: *Tuning of the hyperparameters in the SVM model in Approach 3 while using 100% of the data. Green represents a higher likelihood on the validation data while red represents a higher one. The scale is log_2 for both C and σ^2.*

6.2.2 Tuning of SVR in Approach 3

Since the SVR models uses only observations where the LGDs are non-zero, the run time was much faster. This made it feasible to use the whole data for all choices of hyperparameters to be investigated. In Figure 14, the values of RMSE on the calibration set for hyperparameters in the region $(C, \sigma^2) \in \{2^i, 2^j\}_{(i,j)=(-7,-7)}^{(12,12)}$ are shown. The best fit was found at $(C, \sigma^2) = (2^3, 2^5)$ and these values were therefore chosen.

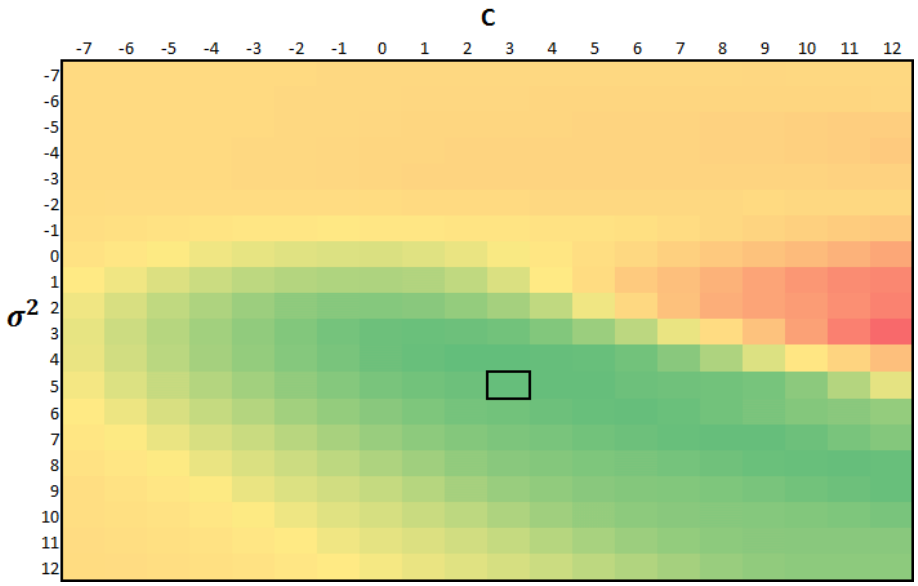

Figure 14: *Tuning of the hyperparameters in the SVR model in Approach 3. Green represents a lower RMSE on the validation data while red represents a higher one. The scale is log_2 for both C and σ^2.*

6.2.3 Tuning of SVM in Approach 4

The SVM model in Approach 4 was calibrated with half of the training data and the log-likelihood was evaluated on half of the calibration data, for hyperparameters in the region $(C, \sigma^2) \in \{2^i, 2^j\}_{(i,j)=(-7,-7)}^{(12,12)}$. In Figure 15, the log-likelihood is shown for different hyperparameters. The figure shows that the region with highest log-likelihood for the SVM model in Approach 4 is the region to the bottom right. The model was recalibrated with parameters in this region, which is marked and dotted, on the full data. The result is shown in Figure 16.

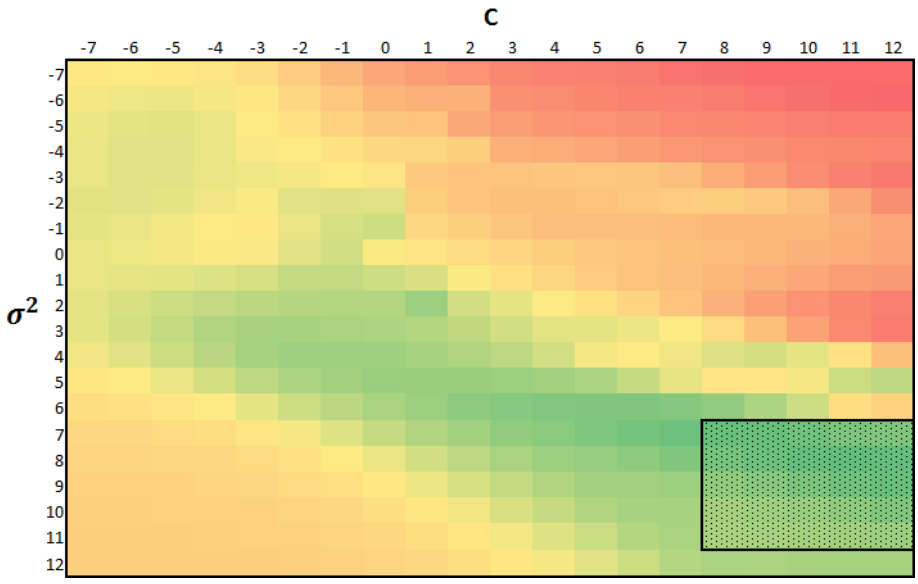

Figure 15: *Tuning of the hyperparameters in the SVM model in Approach 3 while using 50% of the data. Green represents a higher likelihood on the validation data while red represents a higher one. The scale is log_2 for both C and σ^2.*

In Figure 16, the SVM model in Approach 4 has been calibrated with the full data in the region that had the highest likelihood with half of the data, which was shown in Figure 15. The hyperparameters which yielded the best fit was $(C, \sigma^2) = (14, 14)$, which therefore are chosen for the model.

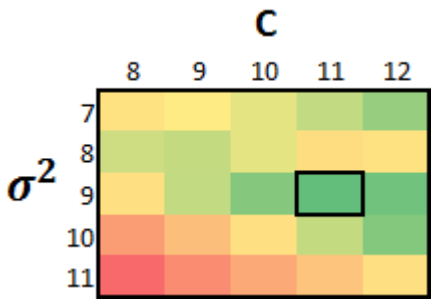

Figure 16: *Tuning of the hyperparameters in the SVM model in Approach 3 while using 50% of the data. Green represents a higher likelihood on the validation data while red represents a higher one. The scale is log_2 for both C and σ^2.*

6.2.4 Tuning of SVR in Approach 4

In Figure 17, the fit in terms of RMSE on the calibration set for different values of the hyperparameters are shown. The best fit was found at $(C, \sigma^2) = (2^{11}, 2^9)$ and these values were therefore chosen.

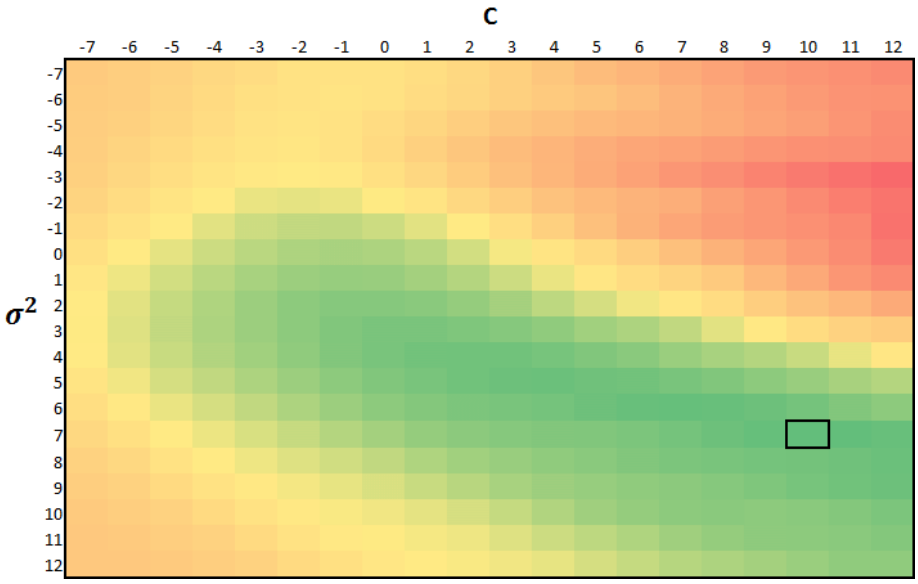

Figure 17: *Tuning of the hyperparameters in the SVR model in Approach 4. Green represents a lower RMSE on the validation data while red represents a higher one. The scale is log_2 for both C and σ^2.*

6.3 Prediction accuracy

Due to data confidentiality issues, on all figures, the numbers on the vertical axis are removed and the curves have been parallel shifted. Furthermore, the RMSEs have been censored as well as Figure 18-20 which show the average levels of LGD over time.

In Table 3, the RMSEs on the data for the total prediction, the loss probability and the conditional severity are shown for all approaches. As can be seen Approach 3 performs worst when it comes to all three measures of RMSE. Approach 1 and 2 are more or less identical, and they fare better than Approach 4 for the conditional severity, while Approach 4 performs best on the loss probability. However in general, all the RMSEs are very similar for the approaches - making it uncertain if there really are any differences.

For setting the RMSE values in a context, the RMSE was in the range 0.2402-0.3258 in the previously mentioned study by X. Yao.[35] In the study by G. Loterman et al the RMSE values differed a lot between the data from the different institutions.[36] The bank with the lowest average RMSE had it differ between 0.1219 and 0.1456 for the models, while the bank with the highest average RMSE had it differ between 0.3299 and 0.3607.

Table 3: *RMSE for the LGD prediction, and for the two underlying models, on the test data for each of the four approaches.*

Approach	**1**	**2**	**3**	**4**
RMSE Loss Given Default				
RMSE Loss Probability				
RMSE Conditional Severity				

In Figure 18 the average realised LGDs are shown for all years in the data and the predicted average value for the different methods are plotted for the last two years, i.e. on the test data that they were not trained on. In Figure 19, the share of the customers with a loss greater than zero is shown, together with the estimated average probability of loss for the different approaches.

[35] Yao et al., 'Support vector regression for loss given default modelling'
[36] Loterman et al., 'Benchmarking regression algorithms for loss given default modeling'

Figure 18: *The figure shows the average realised LGD and the out-of-time predictions for the four approaches for the test data.*

Figure 19: *The figure shows the average realised loss probability and the out-of-time predictions for the four approaches for the test data.*

Figure 20: *The figure shows the average realised conditional severity and the out-of-time predictions for the four approaches for the test data.*

7 Discussion

The prediction accuracy test showed that the RMSEs for the different approaches were surprisingly similar, essentially the same value for the conditional loss severity, and only somewhat better performance with the probability of loss for Approach 4. The result does not seem to support that either of Approach 2 to 4 performed substantially better than Approach 1.

That Approach 2 did not perform better than Approach 1 speaks for that there were not a better capture of time variation in Approach 2, or that the time variation is just a smaller part of what makes up the random component in LGD. Figure 18-20 shows that Approach 2 captures the time variation slightly better than Approach 1 for both the models and the total prediction for both years in the test data.

Based on the performance, Approach 2 seems to have learned something of the time variation with the time varying intercept, but either because of (1) the inherit randomness of the time variation, (2) wrong macroeconomic variable or (3) too short training period, this effect is not very large. All three of these factors probably play their part, but my guess would be that (1) and (3) are the most important. GDP is a very general variable highly correlated with most other economic indicators so even if unemployment had been a better macroeconomic variable, GDP would still capture a large portion of the impact from macro economic variables.

During the time period, the world has seen large changes in the state of the economy and regulatory requirement, which also affects banking practise making the inherit randomness large. Furthermore, eight years are not a very long period, since there has not even been a full business cycle in Sweden during that period. When modelling financial time series, we often find ourselves in the situation to be damned if we do, damned if we do not - longer time series are needed for stable estimates, but older data quickly loses their relevance.

The conclusion regarding Approach 2 is that it has an slight improvement to Approach 1, but my assessment is that the data do not give the model its full justice, due to the relatively low dependency of GDP compared to random time variation movement and noise in the data.

Approach 3 was a failure from a time series dimensional point of view, see Figure 18-20. Even though the prediction for 2014 was more accurate than those of the other approaches, the prediction for 2015 wandered off far from the true value. This indicates that the approach has not actually learned the real relationship between time and the data. It is almost always dubious to apply a model on data where the range of the explanatory variables is disjoint from the range in the training data which proved to be the case here.

Compared to Approach 3, Approach 4 was much better. It seems to actually have learned from the data and predicted the change in average LGD in 2014

very close to the truth, and the change in 2015 was also predicted well. This, combined with the fact that it had the best RMSEs values, makes it the approach with highest performance. However not including the time series in any way could be problematic since it would for a longer time series give equal weight to the year prior as to the year a decade before.

Treating an underlying time variation that cannot be easily explained is hard. Maybe a better approach than the ones employed would be one that combines Approach 2 and 4, i.e. having SVM and SVR models but with a time varying bias. This could in theory combine the best of both models by incorporating the better accuracy of the SVM and SVR models into a model that is designed for the purpose of being point-in-time and forward looking. It would be non-trivial how to fit such a model, but some iterative technique similar to Expectation-Maximisation would probably be an alternative.

The requirement of IFRS 9 that the level of provisions should be forward looking and point-in-time are very hard to fulfil, especially for LGD. In theory it is easy to say that the levels of collective provisions should go up in a severe economic downturn but since data for LGD are available for most banks only one or at most two decades back, it is hard to see clear patterns of the relationship with macro economic variables, and even though it is possible to see a relationship, it is much harder to assess the details of it. However as was shown in this thesis, it is possible with approaches better than a standard one. More research, and especially towards models designed for capturing the time dependency is needed to improve the way that the prediction of LGD can be forward looking and point-in-time.

References

Allen, Linda, DeLong, Gayle and Saunders, Anthony, 'Issues in the credit risk modeling of retail markets', *Journal of Banking & Finance* 28:4 (2004), 727–752.

Basel, *Principles for the management of credit risk* (Bank for International Settlements, 2000).

Bellotti, Tony and Crook, Jonathan, 'Loss given default models incorporating macroeconomic variables for credit cards', *International Journal of Forecasting* 28:1 (2012), 171–182.

Breusch, T.S. and Pagan, A.R., 'A simple test for Heteroscedasticity and Random Coefficient Variation' (1979).

Burnham, Alison J, MacGregor, John F and Viveros, Roman, 'Latent variable multivariate regression modeling', *Chemometrics and Intelligent Laboratory Systems* 48:2 (1999), 167–180.

Frontczak, Robert and Rostek, Stefan, 'Modeling loss given default with stochastic collateral', *Economic Modelling* 44 (2015), 162–170.

Gupton, Greg M et al., 'LOSSCALCTM: Model for predicting loss given default (LGD)', *Moody's KMV, New York* (2002).

Haykin, Simon S et al., *Neural networks and learning machines*, volume 3 (Pearson Upper Saddle River, NJ, USA:, 2009).

Held, L. and Sabanés Bové, D., *Applied Statistical Inference* (Heidelberg, Berlin: Springer, 2014).

Hlawatsch, Stefan and Ostrowski, Sebastian, 'Simulation and estimation of loss given default', *The Journal of Credit Risk* 7:3 (2011), 39.

Hsu, Chih-Wei et al., 'A practical guide to support vector classification' (2003).

Johnston Ross, Emily B and Shibut, Lynn, 'What Drives Loss Given Default? Evidence from Commercial Real Estate Loans at Failed Banks' (2015).

Loterman, Gert et al., 'Benchmarking regression algorithms for loss given default modeling', *International Journal of Forecasting* 28:1 (2012), 161–170.

Platt, John, 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods', *Advances in large margin classifiers* 10:3 (1999), 61–74.

Smola, Alex J and Schölkopf, Bernhard, 'A tutorial on support vector regression', *Statistics and computing* 14:3 (2004), 199–222.

Sundberg, Rolf, *Lineära Statistiska Modeller* (Department of Mathematics, Stockholm University, 2015).

Sundberg, Rolf, *Statistical Modelling by Exponential Families* (Department of Mathematics, Stockholm University, 2016).

Yao, Xiao, Crook, Jonathan and Andreeva, Galina, 'Support vector regression for loss given default modelling', *European Journal of Operational Research* 240:2 (2015), 528–538.

Zhang, Jie and Thomas, Lyn C, 'Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD', *International Journal of Forecasting* 28:1 (2012), 204–215.