UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
School of Mathematical Sciences

Modelling Credit Card Exposure at Default: Novel Applications of Additive Models and Copula Regression

by

Suttisak Wattanawongwan
Master of Science in Financial Mathematics
ORCiD: 0000-0002-4484-8025

A thesis for the degree of
Doctor of Philosophy

July 2021

University of Southampton

Abstract

Faculty of Social Sciences
School of Mathematical Sciences

Doctor of Philosophy

Modelling Credit Card Exposure at Default: Novel Applications of Additive Models and Copula Regression

by Suttisak Wattanawongwan

The thesis comprises three papers that contribute to the consumer credit risk literature by studying the Exposure At Default (EAD) of credit card portfolios. Three novel EAD modelling approaches are proposed, each tackling different practical prediction and interpretation challenges.

The first paper distinguishes between two groups of card borrowers — those whose balance hits the limit as they approach default time, and those who do not. We conjecture that the level of EAD as well as its risk drivers could be significantly different between the two groups. Hence, we propose a two-component mixture model that conditions EAD on these two respective scenarios, using the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework. Having fitted our proposed model to a real-life dataset of credit card defaults, we find that the mean and dispersion of EAD in the two respective submodels are indeed impacted by different risk factors. More importantly, we find that the proposed model produces a clear improvement in predictive performance.

The second paper studies the dependence between the Probability of Default (PD) and credit card balance, and investigates how this dependence impacts EAD and, thus, expected loss estimation. A joint model for PD and balance is introduced by applying the bivariate Copula Generalised Additive Models for Location, Scale and Shape framework. Using this framework, the two responses can be modelled flexibly under the GAMLSS setting while their association can be captured by a suitable copula. The proposed method also addresses potential sample selection bias by extending the analysis to outstanding balance (rather than simply balance at default time, or EAD) over a sample of both defaults and non-defaults. The proposed model is shown to produce more accurate and sufficiently conservative expected loss estimates, at both individual account and portfolio level.

Most EAD modelling research thus far has focused on point estimation approaches, whilst information on extreme quantiles, rather than the mean, can have greater implications in practice. In order to produce conditional quantiles and interval estimates for EAD, the third paper proposes the use of vine copula-based quantile regression. The proposed method automatically avoids the quantile crossing and multicollinearity problems associated with conventional quantile regression and allows relationships between all of the variables of interest (including EAD) to be modelled through a series of pair-copulas. The analysis shows that the proposed model provides better point and interval EAD estimates and more accurately reflects its actual distribution compared to other models.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:..... Date:...July 2021...

Acknowledgements

I would like to thank my fantastic supervisory team, Prof. Christophe Mues, Dr. Ramin Okhrati and Prof. Taufiq Choudhry, who have educated and, at the same time, encouraged me throughout the PhD life. This thesis could not be completed without their invaluable support. I would like to thank Dr. Mee Chi So for her assistance with data preparation and interpretation. I also thank Prof. Tapas Mishra and Dr. Helen Ogden for their valuable feedback on the work and acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

I would also like to thank all relevant officers of the Development and Promotion of Science and Technology (DPST) and the Royal Thai Government Scholarship who have looked after me, both mentally and financially. In addition, I appreciate Ajarn Panumas for her inspirational guidance which escorted me to the PhD path.

I wish to express my gratitude to my friends and colleagues during the study life in the University of Southampton, specially 19 Sherborne family (P'Boo, P'Tam, P'Pim, P'Sandy, P'Fern, P'Nut, Dao, Turk, Min and Kasia), 10039 Maths office (Nikolas, Steve, Rong, P'Oh, David, Xiaohang and Maria), and Suankularb best friends (Popojung, Arm, Mod, Fame and Worawoo).

Lastly, I wish to express my entirely sincere appreciation to my beloved mother, Momnok, and my family (Amah, Jae'Ning, Auntcat, Papa and Sam) whom I have missed and cared the most. They are the world to me. My honourable mention is also given to Chris, who has been always in our memory, for the stories we had weaved together over a six-year period.

Definitions and Abbreviations

CCF	Credit Conversion Factor
CGAMLSS	Bivariate Copula Generalised Additive Models for Location, Scale and Shape
EAD	Exposure At Default
EL	Expected Loss
GAM	Generalised Additive Models
GAMLSS	Generalised Additive Models for Location, Scale and Shape
GLM	Generalised Linear Models
IRB	Internal Ratings-Based approach
LGD	Loss Given Default
OLS	Ordinary Least Squares
PD	Probability of Default
UL	Unexpected Loss
VaR	Value-at-Risk

Chapter 1

Introduction

The necessity of credit risk modelling is well recognised, especially after the global financial crisis in 2007 and 2008 and its consequences for banks and financial institutions. Three credit risk parameters, namely Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD), are required for the calculation of Basel II and III's regulatory capital requirement, which specifies the minimum level of capital that banks must hold. In retail credit risk, EAD has received far less attention than PD and LGD, although its estimation provides benefits to banks beyond the regulatory setting. For example, economic capital, derived in part from EAD, is required to protect the bank and its clients against severe unexpected events (Leow and Crook, 2016). Moreover, obtaining unbiased EAD estimates is beneficial for managing credit limits and risk-based pricing (Gürtler et al., 2018). This thesis, hence, aims to advance the state of the art in EAD modelling.

Considering that corporate credit has thus far received the bulk of the attention in the EAD literature (Gürtler et al., 2018), this thesis will, instead, focus on retail credit, more specifically, credit cards. These form the largest proportion of revolving retail credit for most Advanced Internal Ratings-Based (A-IRB) banks and contribute the largest number of defaults (Qi, 2009). This should enable sufficiently large information for statistical modelling.

Whilst EAD modelling is fairly straightforward for instalment loans, it is challenging for revolving credit because the latter allows customers to draw up to some agreed limit and repay any amount at any time (as long as the minimum payment is met). In order to model the ensuing EAD, the Basel II and III Accords (BCBS, 2017) have implicitly suggested predicting the Credit Conversion Factor, CCF, which is the proportion of the undrawn amount that will be drawn at default time. Despite its popularity, this approach has several drawbacks (Tong et al., 2016). For example, the CCF distribution is highly bimodal and, hence, difficult to model. In light of these downsides, alternative methods have been put forward, including modelling EAD

directly, as a monetary amount (as opposed to a ratio). This thesis adopts the latter strategy and focuses on EAD modelling directly, rather than targeting the CCF level.

The thesis is comprised of three papers, each proposing novel methods to tackle different EAD modelling aspects and testing them on real-life credit card data.

In the first paper (Chapter 2), we consider two distinct groups of credit card borrowers — those who hit the credit limit (i.e. "maxed out" their cards) prior to default and those who did not —, and propose a two-component mixture model. We conjecture that not just the EAD but also its risk drivers could differ substantially between the two groups. The proposed model is developed under the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017), which offers a flexible regression approach that does not restrict EAD to the exponential family and allows its parameters (location, scale and shape) to be modelled as a non-parametric function of the explanatory variables.

In the second paper (Chapter 3), we study the dependence between PD and credit card balance, and how this impacts the EAD estimation and, hence, the expected loss. We introduce the copula approach as a means to capture such dependence at the individual account level, using a joint distribution with marginal GAMLSS models. Hence, PD and EAD can be modelled flexibly and simultaneously with their dependence structure selected from a rich variety of parametric copula functions.

In the third paper (Chapter 4), we study not only estimation for the mean, but, by employing quantile regression, at different quantile levels of EAD, seeing that extreme quantiles can have greater implications in practice. Thus, the whole EAD distribution and interval estimates can be obtained. We avoid the common limitations of quantile crossing and multicollinearity inherent to conventional quantile regression, by introducing the recent development of vine copula-based quantile regression (Kraus and Czado, 2017). This allows modelling the interrelationships between all variables (including EAD) via a series of pair-copulas.

The final chapter (Chapter 5) will conclude by discussing the main contributions made by the thesis and listing some suggestions for further research. The remaining sections of this introduction chapter (Chapter 1) will provide further background on the Basel regulatory credit risk framework, particularly in relation to the risk parameter of EAD. This is then followed by an introduction to two groups of methods used in the papers, namely the GAMLSS framework and copulas. Next, common performance measures used to evaluate credit risk models (and employed in the subsequent papers) are explained. An overview and the main results of the three papers are provided at the end of the chapter.

1.1 The Basel Accords and the Internal Ratings-Based (IRB) approach

While banks provide several types of financial instruments and services to customers and companies, lending is one of their key activities. In so doing, banks incur potential losses if the borrowers they lend to fail to meet their payment obligations. This uncertainty of repayment is what constitutes credit risk. Managing this source of risk has drawn continued interest, both from the perspectives of management and regulation, especially following the global financial crisis of 2007 and 2008. Therefore, a series of statistical models have been developed in order to quantify various aspects of credit risk. Lam (2014) proposed seven dimensions of risk quantification: probability (what is the chance of the event to occur?), exposure (what is the total possible loss from the event?), severity (how much loss is likely to be suffered?), volatility (how unpredictable is the future?), time horizon (how long will the bank be exposed to the risk?), correlation (what is the relation between individual risks?), and capital (how much capital should the bank hold to cover unexpected loss?).

In the wake of several international bank failures, the Basel Committee on Banking Supervision (BCBS) was founded by the central banks of ten countries, at the end of 1974, with the aim of improving financial stability and the quality of banking supervision standards. Two important sets of guidelines on capital adequacy put forward by the Committee are commonly known as the Basel II and Basel III Accords. These have established the international standard for regulation (BCBS, 2017), outlining a risk management control framework for banks. They have set out a risk-sensitive "regulatory capital requirement" which stipulates a minimum level of capital that banks must hold to remain solvent in the face of increased loan defaults. The higher the credit risk faced by the bank, the more capital it must hold. The typical cost of doing business can be seen as expected loss and can be estimated in advance. However, realised loss is usually uncertain and may, in some time periods, significantly exceed the expected loss. The loss under such an adverse scenario is called "unexpected loss", and the Basel accords require that banks hold sufficient capital to absorb it. Figure 1.1 illustrates the concepts of realised, expected, and unexpected loss.

Setting aside too much capital for unexpected loss protection might be suboptimal because such resources could no longer be utilised by banks for generating profits. So, one needs to decide what is the level of conservatism required and set the capital level accordingly. This level of conservatism can be specified in terms of the quantile of a loss distribution (termed Value-at-Risk, or VaR). In Figure 1.2, a VaR set to the 99% quantile of the loss distribution would thus imply that the realised loss would exceed that value with a probability of 0.01; or equivalently, there is a one in one hundred years expectation of a shortfall. Unexpected losses beyond this point can lead to the

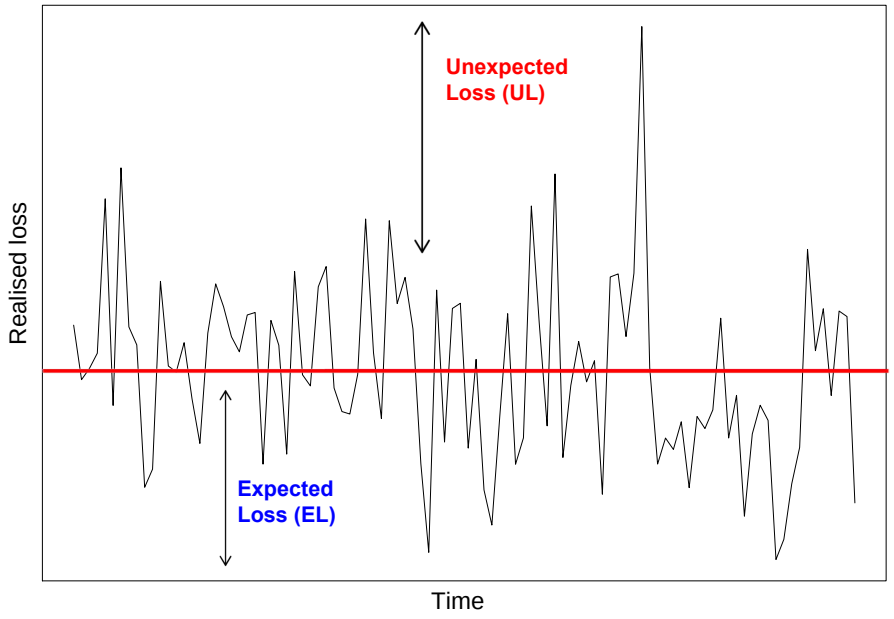


FIGURE 1.1: Example of credit loss dynamic.

bank becoming insolvent. For credit risk, the regulators prescribe a (risk-averse) level of 99.9%, and hence the unexpected loss (to be covered by regulatory capital) is simply the difference between the 99.9% VaR and expected loss.

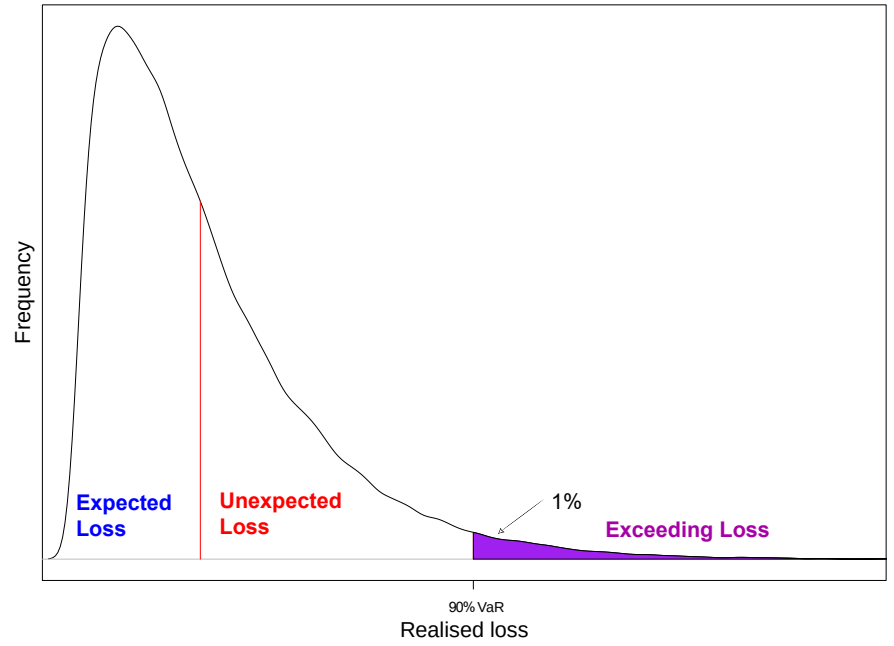


FIGURE 1.2: Example of credit loss distribution.

For credit risk, expected loss is defined (Apostolik et al., 2009) as the product of three key parameters: Probability of Default (PD), i.e. the likelihood that a borrower will default or, in other words, be unable to fulfil their repayment obligations; Exposure At Default (EAD), i.e. the expected gross exposure at the time of default; and Loss Given Default (LGD), i.e. the percentage of this amount that the lender would be unable to recover.

$$\text{Expected Loss (EL)} = \text{PD} \times \text{LGD} \times \text{EAD}.$$

To convert this into unexpected loss, the Accords adopt a version of the so-called Merton model. Merton (1974) proposed the idea of obligors (borrowers) defaulting if their asset value drops below a critical threshold. This model is then extended by assuming that obligors' assets follow a factor model with a common (systematic) factor. This leads to:

$$Y_i = \epsilon_i \sqrt{1-\text{R}} - X\sqrt{\text{R}},$$

where Y_i is the asset value for obligor i, ϵ_i is the idiosyncratic factor (individual risk) for obligor i, X is the common (systematic) factor affecting all obligors, and R is the common asset correlation between obligors. ϵ_i and X are now assumed to be standard normally distributed and independent of each other. Provided the threshold level γ_i, the probability of obligor i defaulting can be expressed as:

$$\text{PD}_i = P(Y_i \leq \gamma_i), \quad \text{leading to} \quad \gamma_i = \Phi^{-1}(\text{PD}_i),$$

where Φ is the standard normal cumulative distribution function. Therefore, the conditional loss given the realisation $X = x$ is:

$$\begin{aligned}
&P\left[Y_i \leq \gamma_i | X = x\right] \times \text{LGD}_i \times \text{EAD}_i \\
&= P\left[\epsilon_i\sqrt{1-\text{R}} - X\sqrt{\text{R}} \leq \Phi^{-1}(\text{PD}_i) | X = x\right] \times \text{LGD}_i \times \text{EAD}_i \\
&= P\left[\epsilon_i \leq \frac{\Phi^{-1}(\text{PD}_i) + x\sqrt{\text{R}}}{\sqrt{1-\text{R}}}\right] \times \text{LGD}_i \times \text{EAD}_i \\
&= \Phi\left[\frac{\Phi^{-1}(\text{PD}_i) + x\sqrt{\text{R}}}{\sqrt{1-\text{R}}}\right] \times \text{LGD}_i \times \text{EAD}_i, \quad \text{since } \epsilon_i \text{ is standard normal.}
\end{aligned}$$

Then, the unexpected loss (UL) is calculated by subtracting the expected loss from the 99.9% VaR:

$$\begin{aligned}
\text{Unexpected Loss}_i &= \text{99.9\% VaR of loss}_i - \text{EL}_i. \\
&= \Phi\left(\frac{\Phi^{-1}(\text{PD}_i) + \Phi^{-1}(0.999)\sqrt{\text{R}}}{\sqrt{1-\text{R}}} - \text{PD}_i\right) \times \text{LGD}_i \times \text{EAD}_i.
\end{aligned} \tag{1.1}$$

According to BCBS (2017), the capital requirement is expressed as a percentage of unexpected loss and fundamentally based on Equation (1.1), with different parameter settings and adjustments for different asset classes. For instance, for retail exposures, minimum capital can be directly derived from Equation (1.1), with the correlation R = 0.15 for residential mortgages and R = 0.04 for revolving products. However, for corporate exposures, further adjustments for maturity, which is the duration or contractual period of loans, have to be made to Equation (1.1) before it can be used to calculate the capital requirement.

1.2 The Basel risk parameters

As previously indicated, the Basel risk parameters PD, LGD, and EAD are required for the calculation of the regulatory capital requirement. To comply with the Basel rules, they are to be estimated for different asset classes of exposures, such as corporate, bank or retail. Moreover, depending on the chosen approach, such estimations can be supplied through internal or external rating systems. Under the Standardised approach, risk weights are prescribed that may depend on external ratings assigned to the obligor. This means loans with higher risk (as measured by the credit rating, grading system, or type of loan) will need more capital as a proportion of exposure size. On the other hand, the Internal Ratings-Based (IRB) approach, introduced by Basel II, allows banks to internally assess credit risk and develop their own statistical models (subject to approval from the regulators). It is further sub-classified into the Foundation-IRB (F-IRB) approach, under which only the PD can be estimated by internal bank models, and the Advanced-IRB (A-IRB) approach, where all three parameters can be internally estimated.

Thus, under the A-IRB approach, the ability to build more accurate models for the three parameters has direct benefits for banks and financial organisations. One of several advantages of employing the A-IRB method is that banks can more accurately assess the risk profile of loans, either at the account level or the portfolio level, by utilising their own data and appropriately chosen models. It implies that the required capital will be more risk-sensitive. This benefits banks with large and high-quality grade credit portfolios, because the lower capital amount required can open up other investment opportunities; conversely, it helps ensure that riskier banks are sufficiently capitalised.

The following subsections briefly elaborate on PD, LGD and EAD.

1.2.1 Probability of Default

The Probability of Default defines the probability that a customer will default within a given time period. Basel considers a one year time horizon over which such default may occur. BCBS defines the default event as when either or both of the following events occur:

> The bank considers that the obligor is unlikely to pay its credit obligations to the banking group in full, without recourse by the bank to actions such as realising security (if held);
>
> The obligor is past due more than 90 days on any material credit obligation to the banking group. Overdrafts will be considered as being past due once the customer has breached an advised limit or been advised of a limit smaller than current outstandings. (BCBS, 2017, p. 93)

With the introduction of Basel II and III, focusing only on the discriminatory power of default prediction models, i.e. the ability to accurately risk rank customers, is no longer sufficient. One also needs strong calibration performance, i.e. being able to produce an accurate estimated PD, as this now is an essential part of the capital requirement calculation (Malik and Thomas, 2007). This additional emphasis on model calibration is also extended to LGD and EAD modelling. Estimating PD can be challenging, though, because of the scarcity of observed defaults. For example, bank portfolios with highly rated customers may contain too few defaults to enable fitting statistical models.

A simple and widely used model for credit scoring and PD modelling is the logistic regression method (Thomas et al., 2017). It estimates PD by indicating whether a borrower is likely to default over a specified period. However, the logistic approach does not take data censoring, commonly found in practice, into account. Data is censored when the event of interest does not occur in the observation period, making it unknown whether that event may occur later. Survival analysis has been introduced in order to deal with this issue. It addresses the censoring problem by applying a survival function to censored data. Moreover, with survival analysis, we model not only *if* borrowers will default (over some fixed period), but also *when* they are likely to default, as it provides dynamic estimates for time-to-default. Banasik et al. (1999) found that the Cox semi-parametric model performance is comparable to that of traditional logistic regression, and sometimes even better. Bellotti and Crook (2009) showed that a Cox proportional hazards (PH) model that includes time-varying macroeconomic covariates alongside borrower-specific variables is superior in terms of the accuracy of predicted PD. Similar findings have been reported by Malik and Thomas (2010), who stated that the driving factors behind default are not fully

explained by the behavioural score alone. Lastly, Tong et al. (2012) applied a mixture cure model, which is an extension of the standard survival model, to the area of credit scoring and compared it with the Cox PH model and logistic regression. The model consists of two parts: an incidence model component which captures the probability of being susceptible to default and a latency model component which predicts the dynamic time-to-default, given that a customer is susceptible and will default at some time point in the future. They concluded that the proposed model is competitive with the other two and produces additional insights.

1.2.2 Loss Given Default

Following a default event, it is not always the case that the bank will lose the entire amount of money owed, as the debt could be (partially) recovered through various channels. For instance, with mortgage loans, banks could resort to repossessing and selling the collateral (i.e. the property of the defaulted borrower), to gain compensation. For example, Yang and Tkachenko (2012) defined LGD as follows:

$$\text{LGD} = 1 - \text{recovery rate, where recovery rate} = \frac{\text{Amount recovered}}{\text{Amount outstanding at default}}.$$

Hence, LGD is the percentage of the exposure that the bank will lose after the recovery or collection processes. The amount recovered are the aggregate discounted cashflows obtained during the recovery period after default time. Depending on the type of loan, it may take several years to work out this value. The resulting LGD is usually (but not always) on the unit scale [0,1].

A key challenge of modelling LGD lies in its observed distribution. Tong et al. (2013) stated that this can be either unimodal or bimodal, being peaked at zero (the debt is fully recovered), and/or one (no recovery achieved). Hence, they decided not to directly model mortgage-loan LGD as a rate, but rather estimate the incurred loss amount via the mixed-discrete zero-adjusted gamma distribution. Their proposed model was compared with two common methods for LGD modelling, namely an Ordinary Least Squares (OLS) model applied to a beta transformed response variable, and tobit regression, and gave competitive calibration performance compared to those models. Somers and Whittaker (2007) used quantile regression to model the house value distribution of repossessed properties, which was then used to estimate the loss given default for mortgage loans. Modelling LGD via tobit regression has been suggested by Bellotti and Crook (2012), in order to account for censoring issues and complying with the regulatory [0,1] interval. However, tobit regression assumes normality of the underlying variable. Sigrist and Stahel (2011) further extended the work by permitting LGD to follow a gamma distribution in the tobit framework. They also took the regular occurrence of zero LGDs into account by using a zero-inflated

model. Leow and Mues (2012) estimated the LGD of mortgage loans via a two-stage model comprising a repossession and haircut model component. The former component allows one to estimate the probability that defaulted loans will go into repossession. The latter predicts the difference between sale price following repossession and the market valuation of the collateral property. The proposed model was shown to outperform a typical single-stage LGD model using standard OLS regression, in terms of the coefficient of determination and the LGD distribution produced.

1.2.3 Exposure At Default

EAD is defined as the outstanding debt at default time, measuring the potential loss banks would incur in the absence of any further repayments. The Basel II and III Accords have implicitly suggested estimating the Credit Conversion Factor (CCF), i.e. the proportion of the undrawn amount at the time of estimation that will be drawn by the time of default, to model the EAD of revolving exposures. However, several drawbacks were soon identified. For example, the CCF distribution is highly bimodal, making it difficult to model. Therefore, in the literature, alternative methods have been suggested, including predicting EAD directly (Tong et al., 2016). More detailed insights from the EAD literature can be found in the literature review sections of each of the papers presented in the following three chapters. In the current chapter, we will restrict ourselves to presenting some additional background relating to EAD modelling that is not included in those.

1.2.3.1 Scarcity of EAD studies

In the credit risk area, PD and LGD have thus far been at the centre of attention, whereas EAD has been studied far less, either in an empirical or theoretical setting. A systematic and extensive literature review on EAD was conducted by Gürtler et al. (2018). They found that most studies were based on the CCF model, and that, in general, the actual observed CCF values are between 30% to 60%. This implies that borrowers typically do not fully draw up to their credit limit when they default. Other modelling strategies were also found, including modelling EAD directly or targeting other relevant EAD factors, e.g. the Loan Equivalent Factor (LEQ), i.e. the exposure at default as a percentage of the outstanding balance at the time of estimation, or the Exposure At Default Factor (EADF), i.e. the proportion of the limit at the estimation time that will be drawn by the time of default. More recent developments in EAD modelling can be found in the work by Thackham and Ma (2018), Gibilaro and Mattarocci (2018) and Luo and Murphy (2020).

As elaborated in Gürtler et al. (2018), the number of papers dedicated to EAD is rather scarce, i.e. approximately 20 over the past two decades. This thesis contributes to this small body of literature by proposing three novel approaches to EAD modelling.

1.2.3.2 Race to default

According to their empirical data, Qi (2009) found that in the so-called "race to default", as default time gets nearer, borrowers tend to be more active than lenders; they tend to draw additional money (increasing the CCF), in contrast to lenders who only very infrequently reduce the limit level or sometimes even increase it. Jacobs and Bag (2010) also concluded that EAD could be impacted by the lender's characteristics and actions. For example, EAD benchmarks of large banks, who may operate an advanced early warning system enabling them to identify deteriorating customers and abruptly decrease their limit before they can borrow more, could be very different from those of small banks with no such efficient detection system.

1.2.3.3 Choice of reference date

For defaulted accounts, the actual value of EAD and the default time are directly observable. In contrast, the choice of a suitable time point for *estimating* the EAD or CCF in advance is not obvious and, to some extent, subjective. It has direct implications for the sample data, too, since realised CCFs for defaulted exposures depend on the point in time when the prior drawn amount and limit are observed; this time point is referred as the "reference date". Likewise, it also determines the time point at which the values of the explanatory variables are to be collected. Below, we summarise three practical methods to identify the reference date, along with their positive and negative aspects.

(1) The fixed-horizon method (Moral, 2006) uses a fixed time interval prior to default, typically setting the reference date to one year before default time (shown in Figure 1.3 with T equal to one year). This method thus implicitly assumes that all accounts that are susceptible to default, will default exactly at the end of the fixed one year horizon. This method provides the benefit of greater homogeneity of the observed response variable (EAD or CCF) but has some limitations as well. First, defaulted facilities whose account tenure at the date of default is less than one year cannot be included in the observation. Second, since only information at the default date, and one year prior to that, are used, other relevant information between those two times is ignored. Third, its assumption that default always occurs at the end of the 12 month outcome period can lead to biased estimates (Moral, 2006).

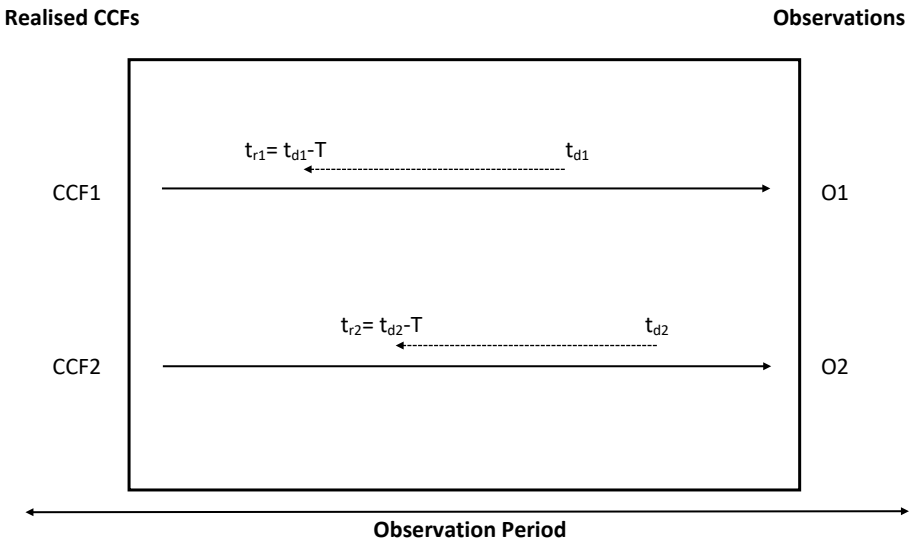


FIGURE 1.3: The fixed-horizon method, where t_{di} is the default date and t_{ri} is the reference date.

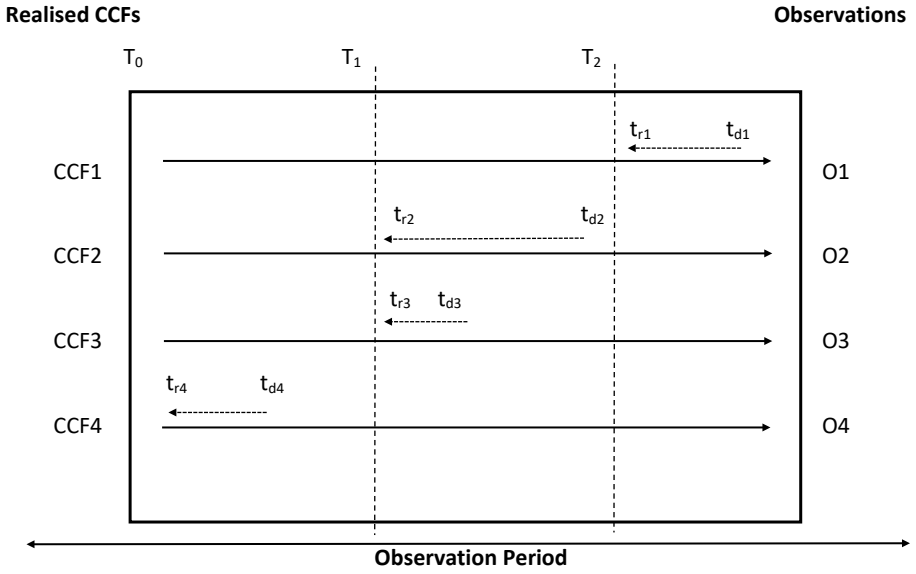


FIGURE 1.4: The cohort method, where t_{di} is the default date and t_{ri} is the reference date.

(2) The cohort method (Moral, 2006) separates the observation period into cohort windows, typically of a one-year length, and sets the reference date to the beginning date of each cohort (shown in Figure 1.4). This reflects a practical implementation by considering that the default could take place at any time point in the following year. It also grants banks the flexibility to select a suitable reference month, avoiding a period where specific circumstances might bias the values of the variables of interest. For example, people tend to use their credit cards more heavily during the winter holiday period; hence, calendar years (starting from January to December) may not be the preferred cohort windows. Similarly to the fixed-horizon method, however, the cohort

method suffers from information loss between the reference and default time. In addition, the realised EAD or CCF values will be less homogeneous than with the fixed-horizon approach, due to the variable time span between reference and default dates (Moral, 2006).

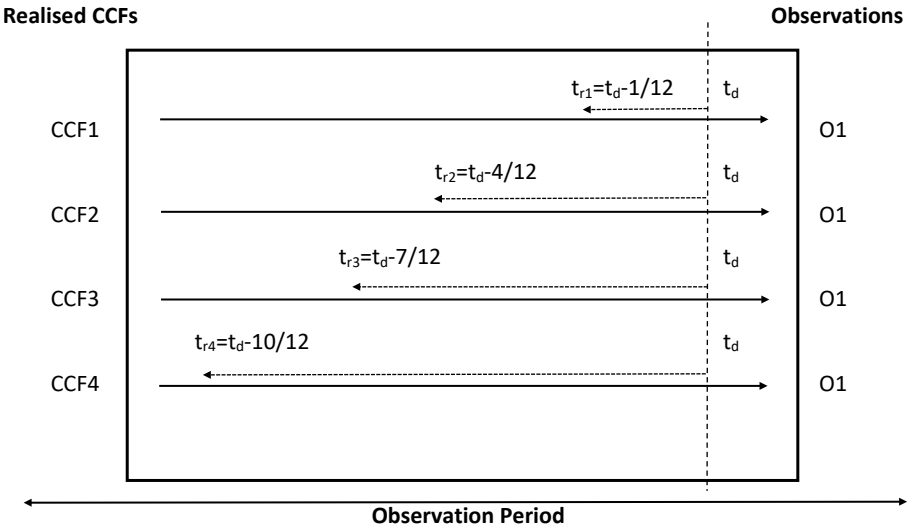


FIGURE 1.5: The variable-horizon method, where t_{di} is the default date and t_{ri} is the reference date.

(3) The variable-horizon method (Witzany, 2011) subdivides the year before default into several time windows with different reference dates (e.g. one to twelve months), see Figure 1.5. Different values of the CCF are thus calculated from the same defaulted exposure, which have to be aggregated later into a single number for EAD estimation. Such a larger number of observations could theoretically lead to more stable and accurate ex-post estimates (Moral, 2006). However, this also means banks need to record a larger amount of data, up to 12 observations per account. Moreover, the homogeneity of estimated CCFs can be questioned.

Although the variable-horizon method might be efficient for the estimation of ex-post CCFs, it does not provide a clear reference date. This poses a problem for producing ex-ante forecasts since the time point where the values of the input variables should be observed is unclear. As the forecast perspective has greater implications for our EAD framework, this suggests that the methods with an explicit reference estimation time, such as fixed-horizon and cohort methods, are preferable.

Among the three approaches, the fixed-horizon approach seems to be the most conservative and is likely to provide higher EAD or CCF estimates than the other methods (Witzany, 2011). However, its implied assumption about the timing of future default events appears less realistic than that of the cohort approach. For this reason,

we will apply the yearly cohort method to prepare the data, and set the reference month to 1st November of each cohort year.

1.2.3.4 PD-weighted estimation of EAD

Witzany (2011) proposed an approach to EAD modelling that involves not only regressing EAD against its drivers but also incorporates default intensity modelling. By separating a one-year time interval into n discrete sequences of subintervals, the CCF can be obtained using a Probability of Default (PD)-weighted approach:

$$CCF = \frac{1}{\sum_{i=1}^{n} \hat{p}_i} \sum_{i=1}^{n} \hat{p}_i \times CCF_i,$$

where CCF_i is the CCF conditioned on default time being in the subinterval $(t_{i-1}, t_i]$ and $\hat{p}_i$ is the probability that default occurs during this interval. In order to obtain each CCF_i estimate, Witzany (2011) recommends using the corresponding fixed time horizon. In the next chapter, we will apply a simplified version of this PD-weighted approach, allowing us to incorporate time to default in our model, as this is shown to have a significant effect on EAD dispersion.

1.2.3.5 EAD modelling challenges

Modelling EAD presents several challenges. First of all, empirical benchmark datasets for EAD have been small in number, or even unavailable (Jacobs and Bag, 2010). Moreover, the range of realised EAD levels could be very wide and its right-skewed and heavy-tailed distribution is difficult to capture statistically (Yang and Tkachenko, 2012). Furthermore, under the A-IRB approach, internal EAD estimates by banks need to meet specific minimum requirements stated by the Basel Accords (Hahn et al., 2011). Some examples of the latter are as follows:

- For on-balance sheet items, the estimated EAD must be no less than the current drawn amount.
- The estimated EAD should reflect the likelihood of additional drawings up to and, possibly, after default occurs.
- EAD must be calculated in a more conservative way if a positive relationship between default frequency and EAD is expected.
- Banks must provide an economic downturn estimate for EAD when EAD varies over the economic cycle.

- The methods for estimating EAD must be practical, intuitive, and reflect what the bank believes to be EAD risk drivers. Also, at least on an annual basis, banks need to review estimation methods, considering newly available information.
- On a daily basis, banks must possess an ability, or systematic processes, to thoroughly monitor outstanding balance changes against the limit level. This enables banks to prevent defaulting borrowers to further draw down money.

The proposed EAD models and data used in this thesis are equipped to cope with the aforementioned issues. For instance, the models are built based on a real-world credit card dataset, comprising more than sixty thousand defaults. This should contribute sufficiently large information about the characteristics of defaulted accounts to enable statistical modelling and avoid the data paucity problem. Moreover, EAD is estimated by a non-parametric approach to quantile regression (see Chapter 4) or a parametric distribution from the flexible GAMLSS framework (see Chapters 2 and 3); hence, any non-standard characteristics of the empirical EAD distribution can be easily captured. Furthermore, the dependence between PD and EAD is taken into account by the copula regression method of Chapter 3, which is shown to produce more conservative EAD and expected loss estimates than when this relationship is neglected. In addition, we focus on building empirical EAD models using account-level covariates, to which macroeconomic variables can be added to reflect EAD in downturn scenarios. Finally, to help ensure that the models are intuitive, we only consider methods that allow us to inspect the effect of each predictor on the EAD target.

1.2.3.6 Revolving credit and credit cards

As noted earlier, EAD modelling is more challenging for some types of credit than for others. For example, the EAD for fixed-term loans, such as residential mortgages and personal loans, can be inferred simply from the current exposure amount plus potential subsequent interest and fees (Witzany, 2011). In contrast, the estimation of EAD for revolving retail exposures, e.g. credit cards and overdrafts, is more complicated, as customers are allowed to draw up to some predetermined limit and repay any amount at any time (as long as the minimum monthly level is met). As a result, each customer's account balance may change substantially in the run-up to default, and using the current balance may severely underestimate the true exposure risk. The estimation could become even more complex when customers move to default abruptly and draw a large amount just before default (Qi, 2009). In this thesis, we develop EAD prediction models for retail credit card portfolios, which has received limited attention compared to the credit risk literature on EAD modelling for corporate customers (Gürtler et al., 2018).

Although, to some, they may simply be a convenient means of payment, credit cards can offer financial flexibility to borrowers who, due to a poor credit rating, do not have access to other credit channels. As a consequence, a credit card borrower with increasing financial difficulties and, thus, a reduced credit score could end up defaulting with an EAD that is substantially higher than the drawn amount at the time of capital calculation (Qi, 2009). Hence, by ignoring the PD (implied by a credit score) when calculating EAD, credit loss might be underestimated. We will tackle this issue later, in Chapter 3, by proposing to jointly model PD, EAD, and their dependence.

1.2.4 PD, LGD, vs. EAD data requirements

In summary, one can perceive PD as the frequency of losses that will occur, while LGD×EAD can be thought of as the size of the actual loss. Within the period of observation, the modelling of PD utilises data from both defaulted and non-defaulted borrowers, whereas, in practice, EAD and LGD models are built using only the accounts that defaulted. Also, where the time period for observing the outcome of interest is concerned, PD and EAD typically imply a one-year time horizon, whereas the work-out period for LGD could last, on average, three to five years due to the long periods of liquidation of defaulted commitments and realisation of collaterals (Hahn et al., 2011).

1.3 GAMLSS framework

In this section, we present background information on the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017), under which several of our EAD models are built. We also elaborate on the advantages it has over its predecessors: linear regression models, Generalised Linear Models (GLMs) (Mccullagh and Nelder, 1989) and Generalised Additive Models (GAMs) (Hastie and Tibshirani, 1986). The GAMLSS models are applied at the account level, which means we model the relationship between the explanatory variables of interest and the response variable and use it to estimate the future outcome for a given account.

1.3.1 Linear regression models

A linear regression model with n data points and p explanatory variables can be written (in matrix form) as:

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is an $n \times 1$ response variable vector, $\mathbf{X}$ is a known $n \times p$ explanatory variable matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of the errors, assumed to be independently identically normally distributed with zero mean and constant variance, i.e. $\boldsymbol{\epsilon} \overset{\text{ind}}{\sim} N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ for an $n \times n$ identity matrix $\boldsymbol{I}_n$. Hence,

$$\mathbf{Y} \overset{\text{ind}}{\sim} N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n), \quad \text{where} \quad \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}.$$

The parameters, $\boldsymbol{\beta}$, are fitted using the Ordinary Least Squares (OLS) method.

This model specification poses some limitations. For instance, the error terms and the response $\mathbf{Y}$ might not always follow a normal distribution for real-world data. Also, the (mean) value of the response could be related to the set of observed predictors in a non-linear way (for example, a change in the values of $\boldsymbol{X}\boldsymbol{\beta}$ might lead to an exponential change in $\mathbf{Y}$). As well, the variance of the errors and the response may not be constant over the predictors' value range. An alternative approach that avoids these limitations is the application of GLMs.

1.3.2 Generalised Linear Models

The framework of GLMs relaxes the normality assumption and allows the response to follow one of several distributions within an exponential family (denoted here as $ExpoFamily$). Also, an invertible monotonic link function, g, is introduced in order to connect the mean parameter $\boldsymbol{\mu}$ to (a linear combination of) the explanatory variables. GLMs can be defined as:

$$\mathbf{Y} \overset{\text{ind}}{\sim} ExpoFamily(\theta, \phi), \quad \text{where} \quad g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta},$$

$\boldsymbol{X}\boldsymbol{\beta}$ is called a linear predictor, and θ and ϕ denote the natural and scale parameters, respectively, of the exponential family. The probability density or mass function of the exponential family is:

$$f_Y(y|\theta, \phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right],$$

in which $a(\phi), b(\theta)$ and $c(y, \phi)$ are some known functions, and $E(Y) = \mu = b'(\theta)$ and $Var(Y) = a(\phi)b''(\theta)$, with $b'(\theta)$ and $b''(\theta)$ denoting the first and second order derivatives of $b(\theta)$, respectively. Note that the variance is no longer constant but depends on the mean level.

GLMs assume that the functional relationship between the response and the explanatory variables can be specified using parametric terms, leading to the

estimation of $\boldsymbol{\beta}$ coefficients. However, the true relationship between the mean and the set of explanatory variables might be more complex, so that the parametric form could not efficiently capture it, resulting in a misspecified model. Hence, GAMs instead suggest the use of smoothing functions to capture such relationships; their idea is to let the data speak for itself and suggest a suitable functional form.

1.3.3 Generalised Additive Models

GAMs can be written as:

$$\mathbf{Y} \overset{\text{ind}}{\sim} ExpoFamily(\boldsymbol{\mu}, \phi), \quad \text{where} \quad g(\boldsymbol{\mu}) = \boldsymbol{X\beta} + s_1(x_1) + \cdots + s_p(x_p),$$

where $x_1, \ldots, x_p$ is a series of explanatory variables and the $s(\cdot)$ represent non-parametric smoothing functions which capture any potential non-linear impact of the predictive variables. The shape of a smooth function is fitted depending on the actual underlying pattern in the data rather than a predetermined set of parameters. Three classes of smoothing techniques commonly used in GAMs are local regression, smoothing splines and regression splines (B-splines, P-splines, thin plate splines). In this thesis, we utilise the penalised B-splines or P-splines because they automatically optimise the trade-off between smoothness and fitness accuracy of the fitted smoothing functions. In GAMs, the word "additive" refers to the fact that in order to evaluate the overall effect of the explanatory variables on the response, we need to add up their individual effects.

One of the problems utilising GAMs, though, is that the variance, skewness, and kurtosis are kept constant for a given mean. The models do not allow one to explicitly model the relationship between any of these three parameters and a set of explanatory variables. Moreover, some empirical distributions may be better represented by distributions other than those of the exponential family. One option to achieve a more flexible model, while still keeping the benefits of GAMs, is to extend it to the GAMLSS framework.

1.3.4 GAMLSS

The Generalised Additive Models for Location, Scale, and Shape (GAMLSS) are a recent class of models defined as:

$$\mathbf{Y} \overset{\text{ind}}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau});$$

$$g^{\mu}(\boldsymbol{\mu}) = \mathbf{X}^{\mu}\boldsymbol{\beta}^{\mu} + s_1^{\mu}(x_1) + \cdots + s_p^{\mu}(x_p);$$

$$g^{\sigma}(\boldsymbol{\sigma}) = \mathbf{X}^{\sigma}\boldsymbol{\beta}^{\sigma} + s_1^{\sigma}(x_1) + \cdots + s_p^{\sigma}(x_p);$$

$$g^{\nu}(\boldsymbol{\nu}) = \mathbf{X}^{\nu}\boldsymbol{\beta}^{\nu} + s_1^{\nu}(x_1) + \cdots + s_p^{\nu}(x_p);$$

$$g^{\tau}(\boldsymbol{\tau}) = \mathbf{X}^{\tau}\boldsymbol{\beta}^{\tau} + s_1^{\tau}(x_1) + \cdots + s_p^{\tau}(x_p),$$

in which $D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ can be chosen from a wide range of distributions that can: (1) be highly skewed or kurtotic; (2) be discrete, continuous or mixed discrete-continuous; (3) exhibit heterogeneity, where the values of the scale and shape parameters vary across predictor levels. The first two parameters (μ and σ) are usually referred to as the location and scale, while the latter two (ν and τ) reflect the shape of the distribution D. The GAMLSS framework is a semi-parametric regression framework because the response is modelled according to a parametric distribution (D), whereas the distributional parameters ($\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$) can be fitted with non-parametric smoothing functions.

In order to understand the main attraction of GAMLSS, Figure 2.1 and Figure 2.5 in Rigby and Stasinopoulos (2010) show the plots of the fitted conditional distribution of y for different values of x for a simple linear model and the GAMLSS framework, respectively. The response distribution at each point of the linear model has a different location (mean), but the same scale (sigma) and shape, as those are fixed. On the other hand, the GAMLSS framework permits the response to have the same type of distribution but with different location, scale, and shape, depending on the level of the predictor. Therefore, GAMLSS is a flexible unifying framework for regression. Another prominent feature is that it offers a good trade-off between predictive accuracy and transparency; models in the GAMLSS class are able to predict the response by means of complex non-parametric structures, but, at the same time, allow the modeller to closely examine the relationship between the response parameters and each of the predictors.

1.3.5 Fitting algorithm for GAMLSS models

In this subsection, we briefly describe the algorithm used to fit the GAMLSS models. Most of the smoothing functions in the GAMLSS framework can be represented as $s(x) = \boldsymbol{Z}\gamma$, where $\boldsymbol{Z}$ is a basis matrix depending on the predictor x, and γ is the vector of parameters to be estimated, subject to a quadratic penalty $\lambda\gamma^T\boldsymbol{G}\gamma$, for a known matrix $\boldsymbol{G} = \boldsymbol{Q}^T\boldsymbol{Q}$ and the smoothing parameter λ (Stasinopoulos et al., 2017). Different smoothing terms have different formulations for $\boldsymbol{Z}$ and $\boldsymbol{Q}$. Hence, the model

can be rewritten as:

$$\mathbf{Y} \overset{\text{ind}}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$g^k(\boldsymbol{k}) = \mathbf{X}^k \boldsymbol{\beta}^k + \sum_{j=1}^{p} \boldsymbol{Z}_j^k \gamma_j^k, \quad \text{where} \quad k = \{\mu, \sigma, \nu, \tau\}.$$

The non-parametric models for parameters μ, σ, ν, τ are fitted by maximising the penalised log likelihood (l_p):

$$l_p = l - \frac{1}{2} \sum_{k=\{\mu,\sigma,\nu,\tau\}} \sum_{j=1}^{p} \lambda_j^k (\gamma_j^k)^T \boldsymbol{G}_j^k \gamma_j^k,$$

where l is the log likelihood function,

$$l = \sum_{i=1}^{n} \log f(y_i | \mu_i, \sigma_i, \nu_i, \tau_i),$$

$f(\cdot)$ is the probability density function of the response distribution D, and n is the number of observations. The parameters that need to be estimated, $\{\boldsymbol{\beta}^k, \gamma_j^k$ and $\lambda_j^k\}$, are fitted by performing the Rigby and Stasinopoulos (RS) algorithm (Stasinopoulos et al., 2017). This fitting algorithm has been proved to be modular and consistently stable for most of the additive terms and distributions, as long as the first and second derivatives of the log likelihood function with respect to the distributional parameters are available (Stasinopoulos et al., 2017).

For a given λ, the RS algorithm generates the estimates of $\boldsymbol{\beta}$ and γ through a series of three nested iterations, the innermost procedure of which is the modified backfitting. More details of the RS algorithm can be found in Stasinopoulos et al. (2017), which involve implementing the Iterative Reweighted Least Squares (IRLS) method repetitively until the global deviance has converged for all three steps. In the RS algorithm, the smoothing terms are modelled by Penalised B-splines (Eilers and Marx, 1996) because they enable the smoothing parameter (λ_j^k) selection to be performed automatically by minimising the Akaike Information Criterion (AIC) = $-2l_p + 2N$, where N is the number of parameters in the model.

1.4 Bivariate Copula GAMLSS

In this section, we present background material relating to copulas and the bivariate Copula Generalised Additive Models for Location, Scale and Shape (referred to from here on as the CGAMLSS) framework (Marra and Radice, 2017a), used in several of our EAD models. Under the CGAMLSS, marginal distribution parameters and their

dependence can be estimated simultaneously, using additive predictor terms at account level.

1.4.1 Copulas

A bivariate cumulative distribution function (CDF), $F_{X,Y}(x, y)$, can be expressed as a combination of two marginal cumulative distribution functions, $F_X(x)$ and $F_Y(y)$, and a dependence structure. The copula function is used to explain how the marginal CDFs are connected. A bivariate copula function, $C_\theta : I^2 \to I$, with $I^2 = [0, 1] \times [0, 1]$ and $I = [0, 1]$ is defined as:

$$C_\theta(u, v) = P(U \leq u, V \leq v), \quad 0 \leq u \leq 1, \quad 0 \leq v \leq 1,$$

where the marginal distributions of U and V are uniform over $[0, 1]$ and θ is a dependence parameter, representing the interaction between the marginals U and V. Thus, a copula function is a joint cumulative distribution function generated from given uniform marginals. Analogously to other CDFs, it shares the common properties of $\lim_{u,v\to-\infty} C_\theta(u, v) = 0$ and $\lim_{u,v\to\infty} C_\theta(u, v) = 1$.

The building block for copulas was introduced by Sklar's theorem (Sklar, 1959). On the one hand, the theorem asserts that there is a copula function C_θ such that for a joint distribution $F_{X,Y}(x, y)$ and marginal CDFs $F_X(x)$ and $F_X(x)$,

$$F_{X,Y}(x, y) = C_\theta(F_X(x), F_Y(y)).$$

If the margins are both continuous, then the copula must be unique. Otherwise, the copula C_θ is unique on the limited domain of $Ran(F_X) \times Ran(F_Y)$, where Ran is the range. Sklar's theorem, thus, illustrates how a copula enables a bivariate response vector to be flexibly constructed by arbitrary marginals and allows their dependence structure to be specified by a suitable choice of copula. On the other hand, the theorem also states that if C_θ is a copula with marginal distributions F_X and F_Y, then $F_{X,Y}(x, y)$ must be a joint distribution.

The joint probability density function, $f_{X,Y}(x, y)$, can be subsequently derived from the joint distribution $F_{X,Y}(x, y)$ in terms of a copula function, as follows:

$$
\begin{aligned}
f_{X,Y}(x,y) &= \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \\
&= \frac{\partial}{\partial y} \cdot \frac{\partial C(F_X(x), F_Y(y))}{\partial x} \\
&= \frac{\partial}{\partial y} \cdot \frac{\partial C(F_X(x), F_Y(y))}{\partial F_X(x)} \cdot \frac{\partial F_X(x)}{\partial x} \\
&= \frac{\partial}{\partial y} \cdot \frac{\partial C(F_X(x), F_Y(y))}{\partial F_X(x)} \cdot f_X(x) \\
&= \frac{\partial}{\partial F_X(x)} \cdot \frac{\partial C(F_X(x), F_Y(y))}{\partial y} \cdot f_X(x) \\
&= \frac{\partial}{\partial F_X(x)} \cdot \frac{\partial C(F_X(x), F_Y(y))}{\partial F_Y(y)} \cdot \frac{\partial F_Y(y)}{\partial y} \cdot f_X(x) \\
&= \frac{\partial^2 C(F_X(x), F_Y(y))}{\partial F_X(x)\, \partial F_Y(y)} \cdot f_Y(y) \cdot f_X(x) \\
&= c(F_X(x), F_Y(y)) \cdot f_Y(y) \cdot f_X(x),
\end{aligned}
\tag{1.2}
$$

where the copula density $c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ is the derivative of a copula function $C(u,v)$ with respect to its marginals. The joint density is, therefore, the product of the two marginal densities and their dependence induced by the copula.

Most copula functions have one or two parameters, (θ, ζ), reflecting the dependence power between the margins. They are different in terms of range and ability to measure several dependence patterns. The latter can be explained by the concept of upper (or right) tail dependence, and lower (or left) tail dependence coefficients, λ_U and λ_L, respectively (Balakrishnan and Lai, 2009):

$$
\begin{aligned}
\lambda_U &= \lim_{u \to 1^-} P(Y > F_Y^{-1}(u) | X > F_X^{-1}(u)); \\
\lambda_L &= \lim_{u \to 0^+} P(Y \leq F_Y^{-1}(u) | X \leq F_X^{-1}(u)).
\end{aligned}
$$

The higher the values of these coefficients, the more concentrated the tail dependence. They can be written as a function of a copula function:

$$
\lambda_U = \lim_{u \to 1^-} \frac{1 - 2u + C(u,u)}{1-u} \quad \text{and} \quad \lambda_L = \lim_{u \to 0^+} \frac{C(u,u)}{u}.
$$

Table 1.1 summarises the range of dependence parameter(s) for each well-known copula, together with their tail dependence structure. For example, the Gaussian and Frank copulas represent a radial symmetry, i.e. a linear correlation. The margins connected by these two copulas share the same level of dependence above or below their means. The strength of dependence is strongest at the centre of the marginal distributions, and gets relatively weak in the tails, as shown from the zero values in the tail dependence coefficients. In contrast, the Clayton copula expresses a correlation intensity at the left tail, whereas its middle and right tail dependencies are weak. It is,

Copula	$C(u,v\|\theta,\zeta)$	Range of (θ,ζ)	λ_L	λ_U
Gaussian	$\Phi_2(\Phi^{-1}(u),\Phi^{-1}(v)\|\theta)$	$\theta \in [-1,1]$	0	0
Frank	$-\frac{1}{\theta}\log\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right]$	$\theta \in \mathbb{R}\backslash\{0\}$	0	0
AMH	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\theta \in [-1,1]$	0	0
Clayton	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$	$\theta \in (0,\infty)$	$2^{-1/\theta}$	-
Joe	$1-[(1-u)^\theta+(1-v)^\theta$ $-(1-u)^\theta(1-v)^\theta]^{1/\theta}$	$\theta \in (1,\infty)$	0	$2-2^{1/\theta}$
Gumbel	$\exp\left(-\left[(-\log u)^\theta+(-\log v)^\theta\right]^{1/\theta}\right)$	$\theta \in [1,\infty)$	0	$2-2^{1/\theta}$
Student-t	$t_{2,\zeta}\left(t_\zeta^{-1}(u),t_\zeta^{-1}(v)\|\theta,\zeta\right)$	$\theta \in [-1,1]$, $\zeta \in (2,\infty)$	$2t_{\zeta+1}\left(-\sqrt{\frac{(1+\zeta)(1-\theta)}{1+\theta}}\right)$	

TABLE 1.1: Commonly used copula functions with the formula specification, the range of dependence parameters (θ,ζ) and their upper, λ_U, and lower, λ_L, tail dependence coefficients. $\Phi_2(\cdot,\cdot|\theta)$ denotes the CDF of the standard bivariate normal distribution with correlation coefficient θ. $\Phi(\cdot)$ denotes the CDF of the standard univariate normal distribution. $t_{2,\zeta}(\cdot,\cdot|\theta,\zeta)$ denotes the CDF of the standard bivariate Student-t distribution with correlation coefficient θ and degree of freedom ζ. $t_\zeta(\cdot)$ denotes the CDF of the standard univariate Student-t distribution.

hence, an appropriate copula for two random variables that exhibit a stronger correlation at low values but weaker at the other areas. Figure 1.6 shows how the dependence parameter θ affects the dependence structure between the two margins. A larger θ lets the Gaussian, Gumbel, and Clayton copulas concentrate more towards their preferences, i.e. at the middle, higher, and lower values of the margins, respectively. Further explanation and theory of copulas can be found in Trivedi and Zimmer (2006) and Nelsen (2006).

Since each copula's dependence parameter(s) has (have) a different range, it is not straightforward to compare the correlation between the margins by using θ or ζ. A more interpretable way is to use a concordance measure, such as Kendall's Tau, τ, which universally falls in the interval $[-1,1]$. A positive (negative) sign and the absolute value in size indicate how strong two variables are positively (negatively) correlated, respectively. The Kendall's Tau can be expressed in terms of a copula function (Balakrishnan and Lai, 2009) as:

$$\tau = 4\int_0^1\int_0^1 C(u,v)c(u,v)\mathrm{dudv} - 1 = 4\mathbb{E}\left[C(U,V)\right] - 1,$$

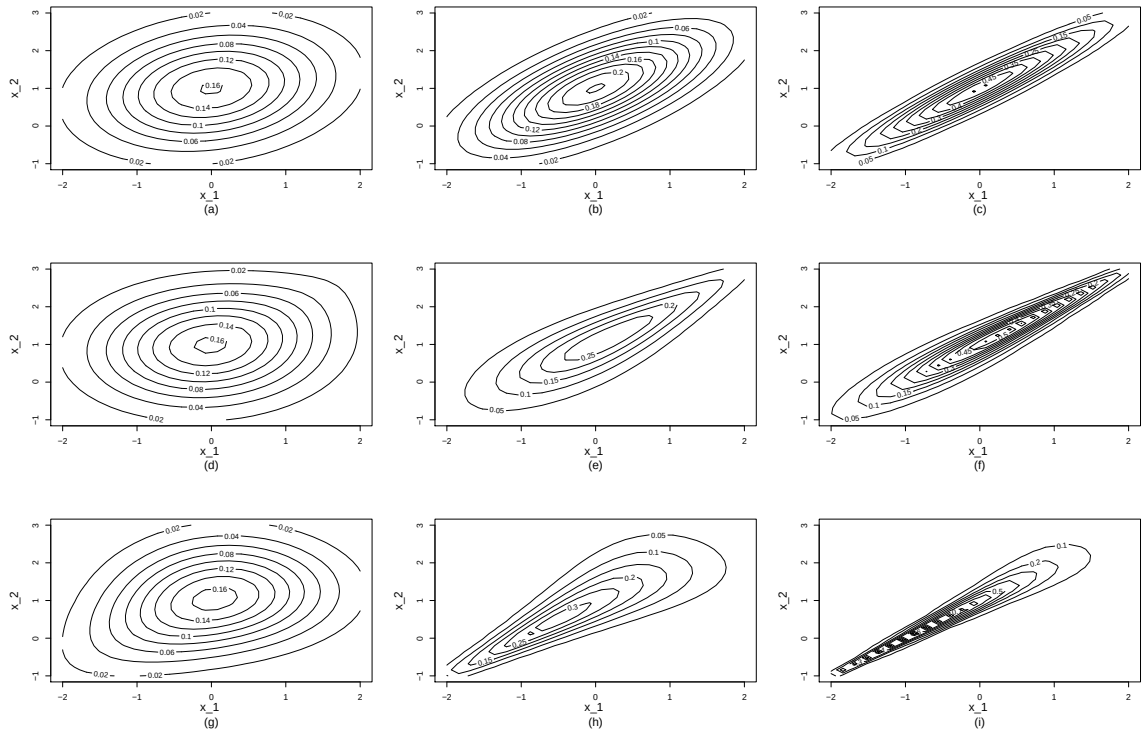


FIGURE 1.6: Contour plots of the Gaussian copula: (a)–(c); the Gumbel copula: (d)–(f); and the Clayton copula: (g)–(i), at different values of dependence parameter θ: (a) $\theta = 0.2$; (b) $\theta = 0.7$; (c) $\theta = 0.95$; (d) $\theta = 1.1$; (e) $\theta = 2.5$; (f) $\theta = 4.8$; (g) $\theta = 0.4$; (h) $\theta = 3$; (i) $\theta = 9$.

where $c(u, v)$ is the the copula density. Table 1.2 shows the relationship between dependence parameters and Kendall's Tau.

Copula	Link function	Kendall's Tau
Gaussian	$\tanh^{-1}(\theta)$	$\frac{2}{\pi}\arcsin(\theta)$
Frank	-	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$
AMH	$\tanh^{-1}(\theta)$	$-\frac{2}{3\theta^2}[\theta + (1-\theta)^2\log(1-\theta)] + 1$
Clayton	$\log(\theta)$	$\frac{\theta}{\theta+2}$
Joe	$\log(\theta - 1)$	$1 + \frac{4}{\theta^2}D_2(\theta)$
Gumbel	$\log(\theta - 1)$	$1 - \frac{1}{\theta}$
Student-t	$\tanh^{-1}(\theta)$, $\log(\zeta - 2)$	$\frac{2}{\pi}\arcsin(\theta)$

TABLE 1.2: Commonly used copula functions with their link functions and the relationship between the respective dependence parameters (θ,ζ) and Kendall's Tau. $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{\exp(t)-1}dt$ and $D_2(\theta) = \int_0^1 t\log(t)(1-t)^{\frac{2(1-\theta)}{\theta}}dt$. Link functions are provided to ensure the appropriate range of the dependence parameters.

Rotated copulas (Nelsen, 2006) have been used to capture asymmetric dependence structures which are not possible with the non-rotated versions. For example, the 90-degree Clayton copula measures a negative correlation, (rather than a positive one as in the original 0-degree Clayton) with an emphasis on the right tail. The 90-degree Joe copula also reflects a negative dependency, but focuses on the left tail. The formulations of the rotated 90-degree (C^{90}), 180-degree (C^{180}) and 270-degree (C^{270}) copulas are:

$$\begin{aligned} C^{90}(u,v) &= v - C(1-u,v), \\ C^{180}(u,v) &= u + v - 1 + C(1-u,1-v), \\ C^{270}(u,v) &= u - C(u,1-v). \end{aligned}$$

1.4.2 Copula GAMLSS

Assuming that the marginal distributions for X and Y have three parameters, namely μ_m, σ_m, ν_m, for $m = 1, 2$, respectively, and a copula C contains two parameters, namely ζ and θ, then a bivariate joint distribution can be expressed as:

$$F_{X,Y}(x, y|\boldsymbol{\vartheta}) = C_{\zeta,\theta}(F_X(x|\mu_1, \sigma_1, \nu_1), F_Y(y|\mu_2, \sigma_2, \nu_2)),$$

where $\boldsymbol{\vartheta} = (\mu_1, \sigma_1, \nu_1, \mu_2, \sigma_2, \nu_2, \zeta, \theta)$ is the vector containing all distributional parameters of the margins and copula function. The bivariate Copula Generalised Additive Models for Location, Scale and Shape (CGAMLSS) framework (Marra and Radice, 2017a) allows a wide range of twice differentiable parametric distributions, with no more than three parameters, to be applied to the margins. The available options for copulas that are extensively used are: the Ali-Mikhail-Haq (AMH), Clayton, Frank, Gaussian, Gumbel, Joe and Student-t, listed in Table 1.2. The link function ensures that the copula parameters lie within their possible range. The augmented copula choices can be selected by rotating the original 0-degree copulas.

Under the CGAMLSS framework, all parameters in $\boldsymbol{\vartheta}$ can be modelled as a function of covariates $\mathbf{z}$, using additive predictors with various effects, such as parametric, non-parametric or splines. Monotonic link functions are used in order to ensure that their parameter space is mapped to the correct range of each distributional parameter. The CGAMLSS model specification is:

$$g_{\mu_1}(\mu_1) = \eta_{\mu_1} = \beta_0^{\mu_1} + \sum_{k=1}^{K_{\mu_1}} s_k^{\mu_1}(\mathbf{z}_k^{\mu_1}); \qquad g_{\mu_2}(\mu_2) = \eta_{\mu_2} = \beta_0^{\mu_2} + \sum_{k=1}^{K_{\mu_2}} s_k^{\mu_2}(\mathbf{z}_k^{\mu_2});$$

$$g_{\sigma_1}(\sigma_1) = \eta_{\sigma_1} = \beta_0^{\sigma_1} + \sum_{k=1}^{K_{\sigma_1}} s_k^{\sigma_1}(\mathbf{z}_k^{\sigma_1}); \qquad g_{\sigma_2}(\sigma_2) = \eta_{\sigma_2} = \beta_0^{\sigma_2} + \sum_{k=1}^{K_{\sigma_2}} s_k^{\sigma_2}(\mathbf{z}_k^{\sigma_2});$$

$$g_{\nu_1}(\nu_1) = \eta_{\nu_1} = \beta_0^{\nu_1} + \sum_{k=1}^{K_{\nu_1}} s_k^{\nu_1}(\mathbf{z}_k^{\nu_1}); \qquad g_{\nu_2}(\nu_2) = \eta_{\nu_2} = \beta_0^{\nu_2} + \sum_{k=1}^{K_{\nu_2}} s_k^{\nu_2}(\mathbf{z}_k^{\nu_2});$$

$$g_\zeta(\zeta) = \eta_\zeta = \beta_0^\zeta + \sum_{k=1}^{K_\zeta} s_k^\zeta(\mathbf{z}_k^\zeta); \qquad g_\theta(\theta) = \eta_\theta = \beta_0^\theta + \sum_{k=1}^{K_\theta} s_k^\theta(\mathbf{z}_k^\theta),$$

where η is a linear predictor inversely linked to its parameter, e.g. $\mu = g_\mu^{-1}(\eta_\mu)$, g is an appropriate monotonic link function, β_0 is an overall intercept, and the K functions $s_k(\mathbf{z}_k)$ are generic effects (linear, non-linear, spatial, etc.), selected depending on the types of covariates $\mathbf{z}_k$. Notice that each parameter may be related to a different series of functions as well as different covariates. Marra and Radice (2017a) showed that each $s_k(\mathbf{z}_k)$ can be represented as:

$$s_k(\mathbf{z}_k) = \sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(\mathbf{z}_k),$$

which is a linear combination of J_k basis functions, $b_{kj_k}(\mathbf{z}_k)$, and coefficients to be estimated, β_{kj_k}. Equivalently, in matrix form:

$$s_k(\mathbf{z}_k) = \boldsymbol{\beta}_k \mathbf{Z}_k,$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k1}, \ldots, \beta_{kJ_k})$ and design matrix $\mathbf{Z}_k = (b_{k1}(z_k), b_{k2}(z_k), \ldots, b_{kJ_k}(z_k))^T$. Hence, the linear predictor can be written as:

$$\eta = \beta_0 + \sum_{k=1}^{K} \boldsymbol{\beta}_k \mathbf{Z}_k.$$

Parameters to be estimated, $\boldsymbol{\beta}_k$, are subjected to a quadratic penalty, $\lambda_k \boldsymbol{\beta}_k^T \mathbf{D}_k \boldsymbol{\beta}_k$, in order to guarantee specific desired properties of the k^{th} function (e.g. smoothness), where λ_k is a smoothing parameter which regulates the shape of the function s and balances the trade-off between accuracy and smoothness, and $\mathbf{D}_k$ is calculated based on the chosen basis functions. Penalty terms can be written in a more compact way as $\boldsymbol{\beta}^T \mathbf{D} \boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$ and $\mathbf{D} = diag(0, \lambda_1 \mathbf{D}_1, \ldots, \lambda_K \mathbf{D}_K)$.

1.4.3 Fitting algorithm for the Copula GAMLSS model

In this subsection, we briefly outline the algorithm used to fit the CGAMLSS model, i.e. the estimation process for the variable coefficients and smoothing parameters. According to Equation (1.2), the log-likelihood function of the CGAMLSS model can be written as:

$$l(\boldsymbol{\delta}) = \sum_{i=1}^{n} \log[c_{\zeta,\theta}(F_X(x|\mu_{1i}, \sigma_{1i}, \nu_{1i}), F_Y(y|\mu_{2i}, \sigma_{2i}, \nu_{2i}))] +$$

$$\sum_{i=1}^{n}(\log[f_X(x|\mu_{1i},\sigma_{1i},\nu_{1i})]+\log[f_Y(y|\mu_{2i},\sigma_{2i},\nu_{2i})]),$$

where $\boldsymbol{\delta}=(\boldsymbol{\beta}_{\mu_1}^T,\boldsymbol{\beta}_{\mu_2}^T,\boldsymbol{\beta}_{\sigma_1}^T,\boldsymbol{\beta}_{\sigma_2}^T,\boldsymbol{\beta}_{\nu_1}^T,\boldsymbol{\beta}_{\nu_2}^T,\boldsymbol{\beta}_{\zeta}^T,\boldsymbol{\beta}_{\theta}^T)^T$, and c is a copula density. The penalised log-likelihood, $l_p(\boldsymbol{\delta})$, is considered:

$$l_p(\boldsymbol{\delta})=l(\boldsymbol{\delta})-\frac{1}{2}\boldsymbol{\delta}^T\mathbf{S}\boldsymbol{\delta},$$

where $\mathbf{S}=diag(\mathbf{D}_{\mu_1},\mathbf{D}_{\mu_2},\mathbf{D}_{\sigma_1},\mathbf{D}_{\sigma_2},\mathbf{D}_{\nu_1},\mathbf{D}_{\nu_2},\mathbf{D}_{\zeta},\mathbf{D}_{\theta})$
and $\boldsymbol{\lambda}=(\boldsymbol{\lambda}_{\mu_1}^T,\boldsymbol{\lambda}_{\mu_2}^T,\boldsymbol{\lambda}_{\sigma_1}^T,\boldsymbol{\lambda}_{\sigma_2}^T,\boldsymbol{\lambda}_{\nu_1}^T,\boldsymbol{\lambda}_{\nu_2}^T,\boldsymbol{\lambda}_{\zeta}^T,\boldsymbol{\lambda}_{\theta}^T)^T$, contained in the $\mathbf{D}$ components, defines the overall smoothing parameter vector. We maximise the penalised log-likelihood by following the methods from Marra and Radice (2017a) who applied the trust region algorithm with integrated automatic multiple smoothing parameter selection. The former is employed for the estimation of the vector $\boldsymbol{\delta}$, whereas the smoothing term ($\boldsymbol{\lambda}$) selection method is the one proposed by Wood (2004). Two fundamental steps are carried out:

Step 1: At iteration a, we keep $\boldsymbol{\lambda}$ fixed, and for given starting values $\boldsymbol{\delta}^{[a]}$, we maximise $l_p(\boldsymbol{\delta})$ using the trust region approach, i.e.

$$\boldsymbol{\delta}^{[a+1]}=\boldsymbol{\delta}^{[a]}+\underset{\mathbf{p}:\|\mathbf{p}\|\leq\Delta^{[a]}}{\arg\min}\breve{l}_p(\boldsymbol{\delta}^{[a]}),$$

where

$$\breve{l}_p=-[l_p+\mathbf{p}^T\mathbf{g}_p^{[a]}+\frac{1}{2}\mathbf{p}^T\mathbf{H}_p^{[a]}\mathbf{p}],$$

and

$$\mathbf{g}_p^{[a]}=\mathbf{g}_p(\boldsymbol{\delta}^{[a]})=\mathbf{g}(\boldsymbol{\delta}^{[a]})-\mathbf{S}\boldsymbol{\delta}^{[a]}\quad\text{and}\quad\mathbf{H}_p^{[a]}=\mathbf{H}_p(\boldsymbol{\delta}^{[a]})=\mathbf{H}(\boldsymbol{\delta}^{[a]})-\mathbf{S}.$$

Here, $\mathbf{g}_p^{[a]}$ and $\mathbf{H}_p^{[a]}$ are used to denote the penalised gradient vector and Hessian matrix, respectively. The former includes $\mathbf{g}_{\mu_1}(\boldsymbol{\delta}^{[a]})=\frac{\partial l(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}_{\mu_1}}|_{\boldsymbol{\beta}_{\mu_1}=\boldsymbol{\beta}_{\mu_1}^{[a]}},\ldots,$
$\mathbf{g}_{\theta}(\boldsymbol{\delta}^{[a]})=\frac{\partial l(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}_{\theta}}|_{\boldsymbol{\beta}_{\theta}=\boldsymbol{\beta}_{\theta}^{[a]}}$, while the latter comprises $\mathbf{H}(\boldsymbol{\delta}^{[a]})_{o,h}=\frac{\partial^2 l(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}_o\partial\boldsymbol{\beta}_h^T}|_{\boldsymbol{\beta}_o=\boldsymbol{\beta}_o^{[a]},\boldsymbol{\beta}_h=\boldsymbol{\beta}_h^{[a]}}$,
where $o,h=\mu_1,\sigma_1,\nu_1,\mu_2,\sigma_2,\nu_2,\zeta,\theta$. Note that $\|\cdot\|$ denotes the Euclidean norm and $\Delta^{[a]}$ the trust region radius, which will be altered over the iterations.

By assuming that the margins are twice differentiable, the estimation procedure can be performed quickly and precisely by using the analytical score and Hessian. For each iteration, the minimiser $\mathbf{p}$ is obtained by applying a quadratic approximation of $-l_p$, constrained within the trust region centred in $\boldsymbol{\delta}^{[a]}$ of radius $\Delta^{[a]}$. It is then used to adjust the radius size for the next iteration (expanding or shrinking the trust region) and decide whether the updated vector $\boldsymbol{\delta}^{[a+1]}$ should be accepted or declined based on

the improvement ratio ϕ (Conn et al., 2000; Nocedal and Wright, 2006):

$$\phi = \frac{l_p(\boldsymbol{\delta}^{[a+1]}) - l_p(\boldsymbol{\delta}^{[a]})}{(\mathbf{g}_p^{[a]})^T\mathbf{p} + \frac{1}{2}\mathbf{p}^T\mathbf{H}_p^{[a]}\mathbf{p}}.$$

If $\frac{1}{4} \leq \phi \leq \frac{3}{4}$, we accept $\boldsymbol{\delta}^{[a+1]}$, and go to the second step. If $\phi < \frac{1}{4}$, we reject $\boldsymbol{\delta}^{[a+1]}$, redefine the radius of next iteration: $\Delta^{[a+1]} = \frac{1}{4} \times \Delta^{[a]}$, and run Step 1 to find $\boldsymbol{\delta}^{[a+2]}$ again by the same approach. Similarly, if $\phi > \frac{3}{4}$, we reject $\boldsymbol{\delta}^{[a+1]}$ and redefine the radius of next iteration: $\Delta^{[a+1]} = 2 \times \Delta^{[a]}$ (Nocedal and Wright, 2006).

The trust region algorithm has been validated as a better approach for this minimisation problem than the line search method (Radice et al., 2015). It is faster and more stable, specifically for non-concave or nearly flat functions.

Step 2: We keep the value of the accepted parameter vector $\boldsymbol{\delta}^{[a+1]}$ fixed, and solve the following problem:

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \|\mathbf{M}^{[a+1]} - \mathbf{A}^{[a+1]}\mathbf{M}^{[a+1]}\|^2 - \overset{\vee}{n} + 2\mathbf{tr}(\mathbf{A}^{[a+1]}),$$

where

$$\mathbf{M}^{[a+1]} = \sqrt{-\mathbf{H}(\boldsymbol{\delta}^{[a+1]})}\boldsymbol{\delta}^{[a+1]} + \sqrt{-\mathbf{H}(\boldsymbol{\delta}^{[a+1]})}^{-1}\mathbf{g}(\boldsymbol{\delta}^{[a+1]}),$$

$$\mathbf{A}^{[a+1]} = \sqrt{-\mathbf{H}(\boldsymbol{\delta}^{[a+1]})}(-\mathbf{H}(\boldsymbol{\delta}^{[a+1]}) + \mathbf{S})^{-1}\sqrt{-\mathbf{H}(\boldsymbol{\delta}^{[a+1]})},$$

where $\mathbf{tr}(\mathbf{A}^{[a+1]})$ denotes the number of effective degrees of freedom of the penalised model, and $\overset{\vee}{n} = 8n$ (if a three-parameter distribution is employed for both margins and the Student-t copula is applied), and n is the sample size. Wood (2004) showed how to solve this problem using the performance iteration idea by Gu (1992). Step 2 is a more convenient and less computationally intensive process since the required gradient vector and Hessian matrix are already derived from the previous step.

The estimations for $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ in Step 1 and 2 are recursively performed until the algorithm satisfies the stopping criterion:

$$\frac{|l(\boldsymbol{\delta}^{[a+1]}) - l(\boldsymbol{\delta}^{[a]})|}{0.1 + |l(\boldsymbol{\delta}^{[a+1]})|} < 1e-07,$$

i.e. until there is no observed improvement in the objective function $l(\boldsymbol{\delta})$. Note that this criterion depends only on the parameter estimates for $\boldsymbol{\delta}$, ignoring the smoothing term $\boldsymbol{\lambda}$. Hence, Marra and Radice (2017a) warned that when smoothing parameters are estimated, the algorithmic convergence is not straightforward to prove and is still an open topic.

The starting values for the marginal parameters $(\boldsymbol{\beta}_{\mu_1}^{[a]}, \boldsymbol{\beta}_{\mu_2}^{[a]}, \boldsymbol{\beta}_{\sigma_1}^{[a]}, \boldsymbol{\beta}_{\sigma_2}^{[a]}, \boldsymbol{\beta}_{\nu_1}^{[a]}, \boldsymbol{\beta}_{\nu_2}^{[a]})$ can be obtained by fitting the GAMLSS univariate models for each margin, as suggested by Marra and Radice (2017a). The resulting GAMLSS coefficient values are subsequently used to initialise the CGAMLSS estimation process. For the copula parameters $(\boldsymbol{\beta}_{\zeta}^{[a]}, \boldsymbol{\beta}_{\theta}^{[a]})$, the initial values can be selected from the empirical Kendall's Tau between the two responses.

In conclusion, the fitting algorithm for the CGAMLSS model proposed by Marra and Radice (2017a) is fast, reliable and easy to implement for several marginal distributions and copulas. The only requirement is the availability of the distributional CDFs and PDFs, along with the derivatives with respect to their parameters.

Similar approaches to CGAMLSS are found in the work by Yee (2016) and Vatter and Chavez-Demoulin (2015). Yee (2016) modelled distributional parameters of a bivariate response with non-linear covariate effects using Vector Generalised Additive Models. However, no automatic way of selecting the best smoothing parameters was suggested, and the number of copula specifications that are available to implement, is small. Vatter and Chavez-Demoulin (2015) estimated distributional parameters of responses' margin and dependence parameter(s) from a copula function separately and independently, by utilising the two-stage technique. According to the simulation study from Marra and Radice (2017a), this two-stage model is less efficient, though, than the CGAMLSS model. Therefore, we propose the CGAMLSS framework as the main approach to model PD, EAD and their dependence structure.

1.5 Performance measures in credit risk modelling

In this section, we present two types of performance metrics commonly applied in the credit risk area: discrimination and calibration measures. Discrimination refers to the ability to accurately risk rank customers and provide an ordinal ranking of the response variable. Calibration considers the accuracy of the model's predictions; that is, how close are they to the actual values? Good risk ranking ability is required, especially for PD, when banks are evaluating the credit risk quality of a customer. Well-calibrated predictions are essential for expected loss or capital requirement calculation purposes, and hence, valuable for all of the Basel risk parameters, i.e. PD, LGD and EAD.

Several measures used in the academic literature and this thesis are described next: the Area Under the ROC curve (AUROC) and Pearson's correlation, which are both measures for discrimination power; and the Hosmer-Lemeshow (HL) Test, the Brier score, the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the Quantile Loss (QL) function, for calibration performance. At the end, we also

introduce the normalised quantile residuals used in the residual plots for checking the model adequacy of GAMLSS models.

1.5.1 AUROC

The Receiver Operating Characteristic (ROC) curve depicts the discrimination ability of a binary response variable, by considering two competing measures: sensitivity and specificity. Sensitivity is also referred to as the "true positive" rate, and is defined as the probability of predicting an observation as positive/success ($\hat{Y}_i = 1$) when the actual observed value is positive/success ($Y_i = 1$). On the other hand, specificity is the "true negative" rate, i.e. the probability of predicting an observation as negative/failure ($\hat{Y}_i = 0$) when the real value of the observation is negative/failure ($Y_i = 0$). The perfect binary model would result in a sensitivity of 100% and a specificity of 100%.

To convert the probabilities produced by a binary model into binary predictions (that then enable us to calculate sensitivity and specificity), we need to set a classification rule via a cut-off point (c). If the predicted probability of an observation is higher than c, it is predicted as positive; otherwise, it is predicted as negative. A higher (lower) value for c implies a smaller (larger) value of sensitivity and a higher (lower) specificity, respectively.

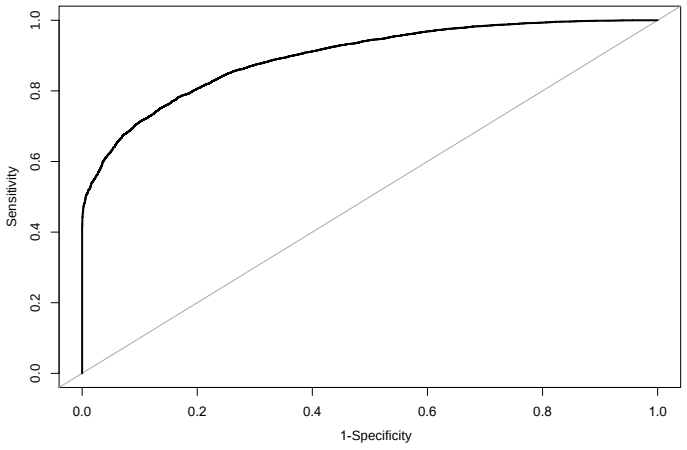


FIGURE 1.7: Example of a ROC curve; the 45-degree diagonal line represents the points where sensitivity = 1-specificity

The ROC curve (see Figure 1.7) plots sensitivity against one minus specificity, as we let the value of c range from zero to one. This plot thus represents the intrinsic trade-off that must be made between sensitivity (ability to detect true positives) and specificity (avoiding false positives). A model with high discriminatory power will have a high value of sensitivity and a low value of one minus specificity; i.e. the ROC curve will stay near the top-left corner of the plot. Conversely, an inefficient model will have the ROC curve near the 45-degree diagonal line. The Area Under the ROC curve

(AUROC) thus provides a measure to quantify and compare the model's quality, varying from 0.5 (worst performance) to one (best performance).

1.5.2 Pearson's correlation

Pearson's correlation coefficient is a statistic measuring the strength of a linear association between two series of values, in this case, predicted and actual values. It is defined as:

$$r_{y,\hat{y}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \quad -1 \leq r_{y,\hat{y}} \leq 1,$$

where n is the sample size, y_i is the actual value, with sample mean $\bar{y}$, and $\hat{y}_i$ is the predicted value, with sample mean $\bar{\hat{y}}$. A positive (negative) value denotes a positive (negative) linear correlation between the variable pair. The higher the absolute value of $r_{y,\hat{y}}$, the stronger the linear association between predicted and actual values, and the better the discrimination performance.

1.5.3 Hosmer-Lemeshow Test

The calibration power of regression models with a binary response can be appraised by the Hosmer-Lemeshow test. For example, a logistic regression model with the response Y and predictors $X_j, j = 1, \ldots, p$ is given by:

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i}.$$

Hence, the predicted probability, $\hat{p}_i$, of success for an account i is

$$\hat{p}_i = P(Y_i = 1) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_p X_{p,i})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_p X_{p,i})},$$

where $\hat{\beta}_j, j = 1, \ldots, p$ are parameters to be estimated.

The Hosmer-Lemeshow (HL) test divides the dataset into g groups based on the value of predicted success probabilities ($\hat{p}_i$), ranking from the lowest to the highest subgroups. For instance, for $g = 10$, the first group contains the observations with the lowest 10% predicted probabilities, followed by the second group consisting of the observations with the next smallest 10%, and so on. Assuming the model is correct, the number of actual accounts with a successful event should match the expected number estimated by the predicted probabilities in each subgroup. The expected number is calculated as the product of the average of predicted probabilities and the number of observations in each subgroup. For example, if the first subgroup has an

average predicted probability of 0.095 and there are 1000 accounts in this subgroup, then we would expect 95 accounts of this first subgroup with a successful event.

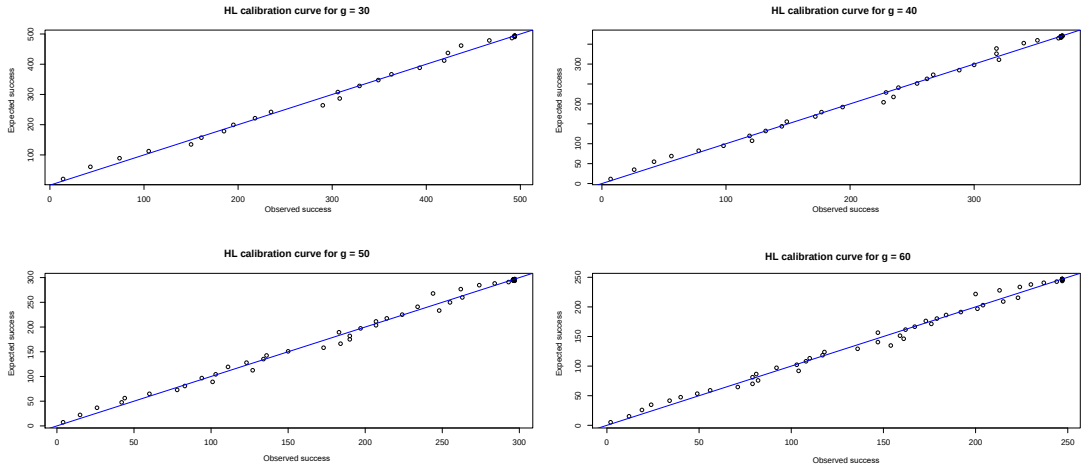


FIGURE 1.8: The Hosmer-Lemeshow calibration curves with different numbers of subgroups g; the 45-degree diagonal line represents the points where the observed number of successes would equal the number of expected successes estimated from the model.

Examples of the calibration curves of the HL test for different numbers of subgroups, g, are displayed in Figure 1.8. The closer the points are to the diagonal line, the better the calibration performance. The choice of the number of subgroups is not uniquely defined. If g is too small, the average predicted probabilities might not reflect the large variation of predicted probabilities in each subgroup, and so the model might still be miscalibrated. On the other hand, if g is too big, the number of observations in each subgroup would be too small, and it would hence be difficult to conclude whether the observed and expected values are different purely by chance or due to a poorly calibrated model. Based on a simulation, Hosmer et al. (2013) suggested the use of $g > p + 1$, where p is the number of covariates in the model.

In order to evaluate the calibration power of the model, the Pearson goodness-of-fit statistic from the HL test is calculated for a partition of g groups, as follows:

$$\sum_{k=0}^{1}\sum_{l=1}^{g}\frac{(O_{kl}-E_{kl})^2}{E_{kl}},$$

where O_{kl} and E_{kl} denote, respectively, the observed and expected number of observations with $Y_1 = k$ in the lth group.

1.5.4 Brier score

The Brier score measures the accuracy of a probabilistic forecast based on the Euclidean distance between the actual outcome and the predicted probability of that

outcome. Low values of the Brier score are desirable, indicating more accurate predictions. The Brier score is defined as:

$$\text{Brier score} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - p_i)^2, \quad 0 \leq \text{Brier score} \leq 1,$$

where n is the sample size, $p_i \in \{0, 1\}$ is the actual outcome, being one when success occurs and zero otherwise, and $\hat{p}_i$ is the predicted success probability of the outcome.

1.5.5 MAE, RMSE and Quantile loss function

The Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) are two of the most common benchmarks used to quantify accuracy for continuous variables. In our context, the errors refer to the difference between actual and predicted values.

MAE is the average magnitude or absolute value of the errors, and is indifferent to the direction of errors. It is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad 0 \leq \text{MAE} \leq \infty,$$

where n is, again, the sample size, y_i is the actual value, and $\hat{y}_i$ is the predicted value. All individual errors have an equal weight of $\frac{1}{n}$.

RMSE also measures the average magnitude of the errors, but using a quadratic rule without again considering their direction. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad 0 \leq \text{RMSE} \leq \infty.$$

Squaring the errors penalises large errors more, and thus the RMSE is appropriate when large mispredictions are particularly unfavourable.

MAE and RMSE apply an equal penalty to overestimation and underestimation. However, there are scenarios where one prefers overestimated rather than underestimated predictions, or vice versa; in such cases, the Quantile Loss (QL) function may be more suitable. QL penalises misestimations differently depending on the choice of a quantile level, α. The α quantile loss function is defined as:

$$\text{QL}(\alpha) = \sum_{i;y_i<\hat{y}_i}(\alpha - 1)\cdot(y_i - \hat{y}_i) + \sum_{i;y_i\geq\hat{y}_i}\alpha\cdot(y_i - \hat{y}_i),$$

where $\alpha \in (0, 1)$. Figure 1.9 considers three different quantile loss functions. When $\alpha = 0.25$, the loss penalty for overestimation is greater. In contrast, the 0.75 quantile loss penalises underestimation more heavily, and, hence, is a good measure for

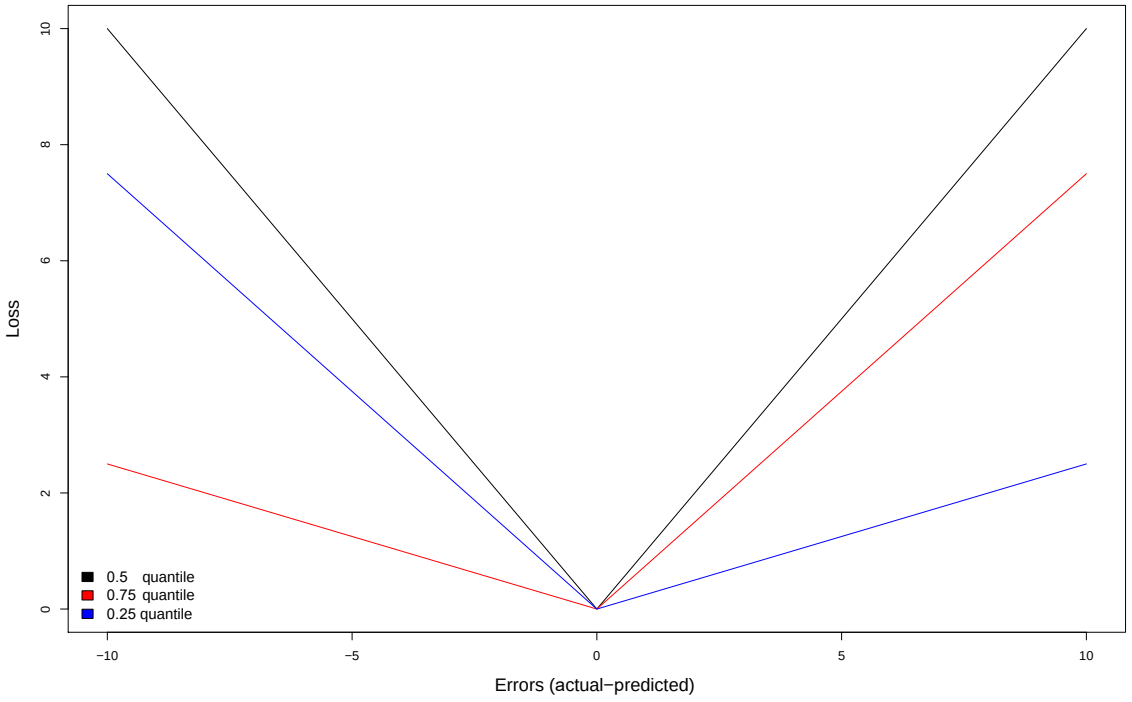


FIGURE 1.9: Quantile loss functions (Y-axis) for different α levels and predicted value (X-axis). The positive and negative error values on the x-axis represent underestimation and overestimation, respectively. MAE is equivalent to the 0.5 quantile loss function.

assessing the conservativeness of risk estimates. For $\alpha = 0.5$, the function returns the MAE.

1.5.6 Normalised quantile residuals

Consider a simple linear regression, defined as $y_i = \beta_0 + \beta_1 X_{1,i} + e_i$, where e_i is the error for an individual i. The raw residuals (ε_i) are interpreted as the difference between the observed and fitted values, i.e. $\varepsilon_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i}$. These raw residuals can be used to check the adequacy of a simple linear model by inspecting whether $\varepsilon_i \sim N(0,1)$. However, in the GAMLSS framework, where the response variable can follow other distributions, it is more sensible to use normalised quantile residuals. Their main advantage is that they always follow a standard normal distribution regardless of the distribution of the response, providing that the model is correctly specified.

Assuming that $F(y|\hat{\theta})$ is the fitted cumulative distribution function with estimated parameters $\hat{\theta}$, the fitted normalised quantile residuals are defined as (Stasinopoulos et al., 2017):

$$\hat{r}_i = \Phi^{-1}[F(y_i|\hat{\theta})],$$

where Φ is the cumulative distribution function of a standard normal distribution. Provided the model is correct, $F(y|\hat{\theta})$ should follow a standard uniform distribution, and hence $\hat{r}_i$ would follow a standard normal distribution. Hence, in order to check the adequacy of the GAMLSS models, regardless of the distribution of the response variable, we could simply check whether the normalised quantile residuals are standard normal.

1.6 Partial residual plots

In the GAMLSS framework, non-parametric smooth functions or splines are commonly used to explain the impact of an explanatory variable on a response, as this impact cannot be explained from an estimated coefficient as with linear regression. In this section, we explain how to interpret the partial residual plots that will be used for the GAMLSS models in the following chapters.

In a simple regression with only one explanatory variable, e.g.

$$y_i = \beta_0 + \beta_1 X_{1,i} + e_i,$$

the value of y_i depends on only one variable $X_{1,i}$, i.e. $\hat{\beta}_1$ shows the effect of $X_{1,i}$ on y_i. However, in multiple regression, e.g.

$$y_i = \gamma_0 + \gamma_1 X_{1,i} + \gamma_2 X_{2,i} + \gamma_3 X_{3,i} + e_i,$$

the value of y_i depends not only on $X_{1,i}$, but also on $X_{2,i}$ and $X_{3,i}$. Hence, $\hat{\gamma}_1$, unlike $\hat{\beta}_1$, tells us the "partial" effect of X_{1i} on y_i, given that all the other explanatory variables are kept constant at a specific value.

A partial residual plot (exemplified in Figure 1.10) is commonly used to identify the nature of the relationship between the target variable and an explanatory variable. This plots the "partial residuals" (on the y-axis) against the respective values of the explanatory variable (on the x-axis). The partial residual for $X_{1,i}$ is defined as the sum of the residual (the difference between observed and fitted value) and the partial effect of the predictor X_1, e.g. $\hat{\gamma}_1 X_{1,i}$ for a linear model. In other words, the partial residual accounts for the part of the response that is not described by the other terms. For a clearer visualisation, the plot depicts a fitted line or smooth function by regressing the predictor's partial residuals against its own value range.

Figure 1.10 shows the estimated non-linear smooth function of the partial residuals of time-to-default on max-out event risk, i.e. the probability that the credit card holder's balance will be at (or above) the credit limit. The y-axis is represented on the logit scale, i.e. $\log\left(\frac{\nu}{1-\nu}\right)$, where ν denotes the probability of a max-out event, since the

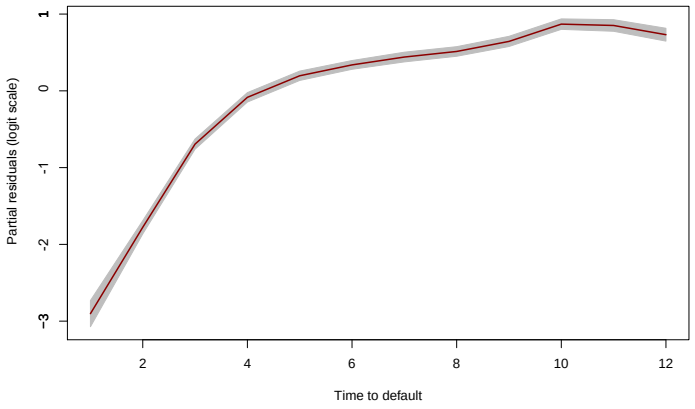


FIGURE 1.10: A partial residual plot of time-to-default (months) for max-out event risk (on logit scale).

logistic additive model was applied. As the time-to-default changes from 5 to 2, the partial residual decreases approximately from 0 to -2. This indicates that, after accounting for the effects of the other explanatory variables, the odds of a max-out event, $\frac{\nu}{1-\nu}$, for a credit card account with a two-month time-to-default are $86\%(1 - e^{-2})$ lower than that with a longer five-month period before default.

1.7 Overview of the three papers

This thesis contains a collection of three papers, each contributing to the literature on EAD modelling for retail credit card portfolios. The proposed models avoid the problems associated with CCF estimation by, instead, modelling EAD directly. Three main methodologies have been employed: the GAMLSS framework, copulas and vine copula-based quantile regression. The ultimate objective of the thesis is to improve the predictive performance of EAD modelling and gain additional interpretation insights, by proposing novel methods and testing them on real-life data.

In the first paper, the newly proposed model combines two ideas found in the literature. First, the model is built under the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework, which produces a much more flexible fitted distribution than its antecedents: Generalised Linear Models (GLMs) and Generalised Additive Models (GAMs). The GAMLSS framework does not restrict EAD to the exponential family and allows for the parameters (location, scale and shape) to be modelled as a function of the explanatory variables. Second, as the level of EAD, as well as the risk drivers for its conditional mean and dispersion parameters, could significantly differ between two subgroups of credit card borrowers — those whose balance hit the limit at least once in the run-up to default, versus those who never maxed out their card over that same outcome period —, we extend our solution to a two-component mixture model, conditioned on these two respective scenarios.

This new model and several benchmark models are empirically validated using a large dataset of credit card defaults from a Hong Kong bank. In addition to identifying the most significant explanatory variables for each model component, our analysis, based on a series of discrimination and calibration criteria, suggests that predictive accuracy is improved by combining the mixture component and the GAMLSS approach.

In the second paper, we introduce a novel approach for modelling PD, EAD, and their dependence structure, simultaneously. The rationale for doing so is that previous studies have shown that expected portfolio loss and, hence, the capital requirement could be underestimated by ignoring the dependence between the credit risk parameters. A joint model for PD and EAD is proposed by applying the bivariate Copula Generalised Additive Models for Location, Scale and Shape (CGAMLSS) framework. Using this model, a joint distribution can be flexibly constructed from two marginal GAMLSS response variables and a suitable copula. Also, whereas most studies have built EAD models on just the subset of defaulted accounts, this new model explicitly addresses potential sample selection bias by extending the analysis to outstanding balance (rather than balance at default time, or equivalently, EAD) over a 12-month period in a sample of both defaulted and non-defaulted accounts. Therefore, this research is the first in the credit risk domain to propose and test a flexible copula regression approach, to simultaneously model PD, card balance and their dependence structure. To empirically validate the effectiveness of introducing the dependence, we also construct two standalone models, for PD and balance separately, against which we benchmark our newly proposed copula model. For a large dataset of credit card accounts, the results reveal strong and positive dependence between PD and balance, even after accounting for observable covariates, either in the middle or upper tail area of the marginal distributions. Moreover, our analysis shows that the proposed model provides more accurate and conservative expected loss estimates, exhibiting a heavy tail that is the result of the correlation between PD and credit card balance. In addition, we demonstrate that by, instead, ignoring such dependence or by allowing sample selection bias, loss could be severely underestimated, potentially leading to capital shortfalls.

In the third paper, the vine copula-based quantile regression model is proposed to estimate conditional mean and quantiles and interval estimates for EAD. This model addresses two key aspects of EAD modelling. First, quantile regression is applied to produce the information on extreme risks in the right tail area, i.e. higher quantiles of EAD, which is useful for risk management and capital calculation. Second, since many of the input variables used in EAD models are strongly correlated with each other (see e.g. current credit limit and balance in Tong et al. (2016) and Leow and Crook (2016)), vine copulas — a flexible class of dependence models — are introduced to model any multi-dimensional dependencies among those variables (including EAD) by a suitable series of (either parametric or non-parametric) pair-copulas. The

vine copula approach avoids two drawbacks of classical quantile regression: the problem of quantile crossing (i.e. the crossing of regression lines of different quantile levels, causing interpretation difficulties) and its difficulty coping with multicollinearity. When tested on a large dataset of credit card defaults, the proposed non-parametric model leads to better point and interval EAD estimates and more closely reflects its actual distribution compared to other models.

1.8 Author contributions

I, Suttisak Wattanawongwan, hereby declare that I am the main author of the three papers who conceived, developed, and implemented the work. The other authors are part of the supervisory team who enlightened and encouraged me and essentially improved the thesis. I also wrote the initial draft of the papers which was further advanced by the supervisory team.

Chapter 2

A Mixture Model for Credit Card Exposure at Default using the GAMLSS Framework

Suttisak Wattanawongwan [a,1], Christophe Mues [b], Ramin Okhrati [c], Taufiq Choudhry [b], Mee Chi So [b]

[a] School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[b] Southampton Business School, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[c] Institute of Finance and Technology, University College London, London, WC1E 6BT, UK

[1] Corresponding author
Email address : S.Wattanawongwan@soton.ac.uk

Abstract

The Basel II and III Accords propose estimating the Credit Conversion Factor (CCF) to model Exposure At Default (EAD) for credit cards and other forms of revolving credit. Alternatively, recent work has suggested it may be beneficial to predict the EAD directly, i.e. modelling the balance as a function of a series of risk drivers. In this paper, we propose a novel approach combining two ideas proposed in the literature and test its effectiveness using a large dataset of credit card defaults not previously used in the EAD literature. We predict EAD by fitting a regression model using the Generalised

Additive Models for Location, Scale and Shape (GAMLSS) framework. We conjecture that EAD level and the risk drivers of its mean and dispersion parameters could substantially differ between the debtors who hit the credit limit (i.e. "maxed out" their cards) prior to default and those who did not, and thus implement a mixture model conditioning on these two respective scenarios. In addition to identifying the most significant explanatory variables for each model component, our analysis suggests that predictive accuracy is improved, both by using GAMLSS (and its ability to incorporate non-linear effects) as well as by introducing the mixture component.

2.1 Introduction

The Basel regulatory accords have set out risk-sensitive regulatory capital requirements stipulating the minimum level of capital that banks must hold as a function of various types of risk. Under the Advanced Internal Rating Based (A-IRB) approach, authorised banks are permitted to use their own methods to calculate three parameters that are central to one such source of risk — credit risk. These are: Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD). In retail credit risk, PD and LGD have thus far received the bulk of attention by credit risk researchers, whilst EAD has been studied far less extensively. This paper is motivated by this fact and aims to close such gap by focusing on EAD modelling.

EAD is defined as the outstanding debt at the time of default and measures the potential loss the bank would face in the absence of any further repayments. The A-IRB approach requires producing suitable EAD estimates for all loans that are not yet in default. For some types of loans, those estimates can be relatively straightforward; for example, the EAD for term loans, such as residential mortgages and personal loans, could be inferred simply from the current exposure amount plus potential subsequent interest and fees (Witzany, 2011). In contrast, for revolving retail exposures, such as credit cards and overdrafts, the estimation is more complex as customers are allowed to draw up to a specified limit and can repay any amount at any time (as long as the minimum level is met). As a result, each borrower's account balance may change substantially in the run-up to default and using the current balance may severely underestimate the true exposure risk. For these types of credit, the Basel Accords have suggested estimating a Credit Conversion Factor (CCF), which is usually defined as the proportion of the undrawn amount (i.e. credit limit minus drawn amount) that will be drawn by the time of default. This CCF should reflect the likelihood of additional drawings between estimation and default time. From the predicted CCF, the estimated EAD then follows as:

$$\text{EAD} = \text{Current drawn amount} + (\text{CCF} \times \text{Current undrawn amount}).$$

Even though statistical methods to estimate the CCF have been proposed, several drawbacks were soon identified. For example, the CCF distribution is highly bimodal, estimates must be restricted to the [0,1] range, and models may struggle to cope with the contracting denominator when the current drawn amount is already close to the limit. Therefore, in the literature, alternative methods have been suggested to avoid the undesired properties of CCF models, which include predicting EAD directly (Tong et al., 2016).

In this paper, we focus on EAD modelling for credit cards, which has received limited attention in the literature. Most of the studies on EAD modelling have thus far focused on corporate credit, whilst fewer address retail customers (Gürtler et al., 2018). This is partly explained by the greater availability of public data on the corporate sector and by the fact that the financial status and health of corporate customers could be inspected from share and market-traded products (Leow and Mues, 2012), enabling easier access to data. Nonetheless, credit cards make up the largest share of revolving retail credit for most A-IRB banks and contribute the largest number of defaults compared to other revolving line products (Qi, 2009). This should contribute sufficiently large information about the characteristics of defaulted accounts to enable statistical modelling.

To avoid the problems associated with CCF estimation, we choose the EAD amount itself as the response variable. This choice, however, poses other challenges. For example, the observed value range of realised EAD levels could be very wide and thus difficult to capture statistically (Yang and Tkachenko, 2012). To cope with its right-skewness, Tong et al. (2016) therefore proposed a gamma distribution for (non-zero) EAD and built a direct EAD model under the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017), which was shown to outperform several benchmark models (including for CCF) on a dataset from a UK lender. In this paper, we take a similar approach but we further extend it by distinguishing between two subgroups of credit card borrowers — those whose balance hit the limit at least once in the run-up to default, versus those who never maxed out their card over that same outcome period —, introducing two mixture components to our models. The rationale for doing so is that we hypothesise that not just the EAD but also its risk drivers (and that of its dispersion) could differ substantially between the two groups. A similar mixture element was previously proposed by Leow and Crook (2016), along with their panel models for card balance (and limit), but besides us using a different modelling framework applied to (cross-sectional) default cohort data, our approach differs from theirs in that we allow for non-parametric terms, and nor do we assume that the balance of maxed-out accounts has to match the limit value exactly.

To empirically validate the effectiveness of the GAMLSS model (versus OLS), the proposed mixture approach, and its combined application, we construct a set of

benchmark models against which we compare the predictive performance of our newly proposed model. All models are fitted using a large dataset of credit card defaults from a Hong Kong lender, which has not been previously used in the EAD literature.

To summarise, the contributions of our new model and analysis are that we: (1) estimate EAD directly, instead of using the conventional CCF approach; (2) analyse EAD in the hitherto underresearched area of retail credit cards; (3) apply the idea of EAD mixture models under the GAMLSS framework and compare its performance to a series of benchmark models; (4) identify the factors that significantly impact the mean and dispersion of EAD, giving further insights into the risk drivers of EAD; (5) inspect any differences in the risk drivers depending on whether the account hit the limit prior to default.

The paper is structured as follows. In Section 2.2, the existing literature on EAD modelling is reviewed. Section 2.3 explains the data and variables used and Section 2.4 illustrates how statistical models are constructed. The results are presented and discussed in Section 2.5. Section 2.6 concludes.

2.2 Literature Review

In order to model the EAD of revolving exposures, the Basel II and III Accords have implicitly suggested estimating a Credit Conversion Factor (CCF), which is the proportion of the undrawn amount at the time of estimation (i.e. credit limit minus current balance) that will be drawn by the time of default, i.e.:

$$CCF_{t,\tau} = \frac{\text{EAD}_{t,\tau} - \text{Balance}_t}{\text{Limit}_t - \text{Balance}_t}.$$

Balance_t denotes the amount of money owed by credit card borrowers at the present time (t). Limit_t is the credit limit or maximum amount that the borrowers could draw at t. $\text{EAD}_{t,\tau}$ is simply the balance at the future default time (τ) estimated at the present time (t). Hence,

$$\text{EAD}_{t,\tau} = \text{Balance}_t + CCF_{t,\tau} \times (\text{Limit}_t - \text{Balance}_t).$$

Analysing CCFs (or other EAD proxies that incorporate current balance and limit) is deemed important because current exposure alone does not give a reliable indication of the final balance at default. The reason is that, as obligors are approaching default, they may draw additional money (or, in some cases, pay back part of the balance). Gürtler et al. (2018) found that the most relevant factors affecting CCF are time to default and borrower risk (credit quality). Moreover, CCF values heavily depend on

the type of product (corporate or retail), data, and empirical methodology used. In the corporate setting, Gibilaro and Mattarocci (2018) also considered the impact of firms having multiple banking relationships, finding that by considering the exposures as a group rather than individually, one could enhance statistical model fit (in terms of R^2) and reduce the risk of underestimation.

CCF distributions tend to be highly bimodal with a probability mass at zero (when there is no change in balance) and another at one (when borrowers end up drawing the entire limit), while showing a flat distribution in between. This causes difficulties in modelling and predictions produced by a conventional Ordinary Least Squares (OLS) regression model could be poor. Therefore, various techniques and models have been put forward as better alternatives for modelling CCF, e.g., Binary logit and Cumulative logit regression models (Brown, 2011), Beta link generalised linear models (Jacobs, 2010), and Naive Bayesian models and single layer neural networks (Yang and Tkachenko, 2012). Empirical evidence suggests most of these produce better performance than OLS regression.

Even though indirect EAD models based on the CCF are commonly used, several other drawbacks have been identified. For example, when the current drawn balance is already close to the limit to begin with, CCF values can become very large and unstable due to the contracting denominator (or even undefined when balance equals limit). This is not uncommon for accounts that will eventually default. Hence, restrictions must be imposed on CCF models (via truncation or censoring), causing loss of potentially useful information. More well-behaved values could be equally problematic, however. For example, Leow and Crook (2016) pointed out that a positive value of CCF can be observed under two different circumstances: (1) when the current balance is less than both balance at default and current limit (which is a common occurrence for accounts going into default); or (2) when the current balance is greater than both balance at default and current limit. Although these two cases may result in the same positive range of CCF values, their characteristics and implications for EAD risk are totally different. This makes the CCF estimate more difficult to interpret. Furthermore, Taplin et al. (2007) illustrated how predicted values greater than one also create undesirable outcomes. Firstly, they would imply that as the balance increases, EAD (and thus the risk) will decrease, which is counter-intuitive because larger balance should intuitively mean larger exposure. Secondly, when the predicted CCF is greater than one and balance is greater than limit, the estimated EAD would be smaller than both balance and limit, which is unlikely to occur. For regulatory capital requirement purposes, the Basel Accords therefore impose calculated CCF values to be strictly in the [0,1] range. However, in real-life datasets, one can often see a large number of CCF observations that are either negative or exceed one. They could be negative when EAD is less than current balance (i.e. the debtor pays back part of the debt before defaulting), providing that balance is below

limit. This more often happens when time to default is large and current credit utilisation is close to one (Moral, 2006). Alternatively, in the empirical dataset analysed by Taplin et al. (2007), 38 percent of all accounts exhibited negative CCF values because they started off with a balance that exceeded the limit (which is contrary to the CCF's core idea of the balance increasing by a fraction of the undrawn amount). Conversely, a sizable proportion of observed CCFs may be greater than one because, in practice, the balance at default time commonly goes beyond the current credit limit, e.g. due to interest and other charges or credit limit increases between t and τ (Tong et al., 2016). Imposing a ceiling on CCF would mean that no EAD estimates could ever exceed the current limit level, which may not reflect reality.

With these obstacles in mind, Luo and Murphy (2020) avoided CCF by implementing other EAD factors, namely EADF ($\text{EAD}_{t,\tau}/\text{Limit}_t$) and AUF ($(\text{EAD}_{t,\tau} - \text{Balance}_t)/\text{Limit}_t$), when estimating EAD in the context of U.S. construction loans. However, these measures might not offer a better alternative. Being a ratio of EAD, the predictor effect on balance upon default cannot be directly obtained. This poses difficulties for practitioners as the interpretation of the relationship between predictors and EAD is important. Also, Leow and Crook (2016) indicated that, as an account approaches default and balance, and hence EAD, increases, lenders act differently; some increase the limit level, some reduce it. This leads to a heterogeneity problem in a cross-sectional model. Moreover, similarly to CCF, the value of EADF is expected to range between zero and one. However, it is not uncommon to perceive outstanding balances go over their limits, resulting in values much greater than one and, therefore, the challenging choice of distribution. Hence, similar restrictions must be imposed, either via truncation or censoring.

In light of these drawbacks, alternative approaches have been proposed that involve modelling EAD directly, as a monetary amount (as opposed to ratio). For example, Thackham and Ma (2018) suggested that, for large corporate revolving facilities, banks often actively manage the borrower's limit amount as default time approaches, and that these changes in limit (up or down) have a large impact on EAD. Therefore, they proposed a mixture (two-stage) model, conditioning their EAD target variable on whether the limit is decreased or not. Hon and Bellotti (2016) did not forecast drawn balance at default time (EAD) as such, but instead proposed models to estimate drawn credit card balance at every time step, unconditional on a default event. They argued that, apart from having risk management applications, the prediction of this unconditional balance on revolving credit lines is beneficial because it provides banks an expected profit estimate. Different models were considered, including OLS, two-stage, mixture regression and random effects panel models. The direct EAD model proposed by Tong et al. (2016) uses a zero-adjusted gamma (ZAGA) distribution to capture the EAD distribution observed in a dataset of credit card defaults, grouped per default cohort. They constructed a model in the GAMLSS

framework, the predictive performance of which they compared against that of three common CCF models and a utilisation change model. The results confirmed that the direct EAD model is a competitive alternative to these benchmark models. Lastly, another mixture model is proposed by Leow and Crook (2016). Using a portfolio of defaulted credit card accounts and their monthly observations, they analysed outstanding balance. Similarly to Hon and Bellotti (2016), they did so not only at the time of default, but at any time over the entire period up to the default time. In addition, they proposed modelling the probability that account borrowing reaches (or exceeds) the limit level at any time period; under that scenario, they proposed modelling the limit rather than the balance. A discrete-time repeated events survival model and panel models with random effects were applied to estimate the former probability and the conditional balance or limit, respectively, which were shown to provide competitive model fit and predictive accuracy compared to conventional models. As with other such panel models, suitable lags would have to be introduced to make the approach suitable to EAD prediction under Basel, which generally assumes a one-year horizon.

Regardless of the method used to model EAD, common major drivers of EAD according to the literature are commitment limit level, current balance, credit utilisation, credit quality, time to default, and undrawn percentage $(1-(\text{Balance}/\text{Limit}))$. In this paper, we use the same variables, supplemented by further behavioural variables derived from monthly account data, as well as a selection of macroeconomic covariates.

Similarly to Tong et al. (2016), our newly proposed direct EAD model is built under the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017). This framework allows selecting a distribution for the response variable, the parameters of which (location, scale, and shape) can be modelled as a function of explanatory variables, either parametrically or non-parametrically. GAMLSS is much more flexible than the Generalised Linear Model (GLM) or Generalised Additive Model (GAM) frameworks, which are restricted to the exponential family. It potentially allows the fitted distribution to (1) be highly skewed and kurtotic, (2) be discrete, continuous, or mixed discrete-continuous, (3) exhibit heteroscedasticity, whereby the value of scale and shape parameters varies across covariate levels. This is important for observed EAD data as it typically exhibits several of these features. Moreover, the ability to model the dispersion of EAD as a function of explanatory variables can be useful from a risk management perspective; where the estimated EAD dispersion is large, we could thus make the point estimate more conservative in order to deal with the greater uncertainty. Motivated by the empirical results reported by Leow and Crook (2016), we further extend the approach by considering that, as accounts move towards default, the balance could either hit the limit or not. This breaks the EAD model into

two mixture components, which could have different EAD levels and risk drivers. Although considering similar scenarios, our approach differs from that taken by Leow and Crook (2016) in a number of ways. First, rather than treating balance as panel data, we apply the default cohort approach in EAD modelling and group defaults according to 12-month calendar periods, as this facilitates producing estimates that are conditional on default and matches the prediction horizon used for Basel. Second, using the GAMLSS framework for all model parts offers a wider range of distributions and, importantly, allows introducing non-linearity. Third, considering that the balance can further vary over time and may exceed the (prior) credit limit, we do not fix the EAD to the credit limit value conditional on a max-out event, but allow its distribution to be explicitly modelled in this mixture component as well, thus giving further insights into specific risk drivers of EAD for this subgroup.

Note that our proposed model is an account-level one; in other words, it is the result of taking a bottom-up approach. More generally, the underlying parameters in credit risk modelling can be estimated in two different ways: top-down or bottom-up (BCBS, 1999). The former approach aggregates data with similar risk profiles, e.g. with regard to credit rating and tenure, and groups them into homogeneous pools, for which well-calibrated credit risk parameter estimates are then provided. This method is typically applied to consumer, credit card or other retail portfolios, due to their volume. For example, Witzany (2011) showed how EAD could be estimated at the aggregated pool level by the top-down approach. On the other hand, the bottom-up approach measures credit risk at an individual (loan or account) level, considering information on the entire set of (inhomogeneous) loans. This approach is often adopted for corporate exposures and capital market instruments. In the consumer credit risk literature, both of these approaches are well known and have each been employed; however, one does not rule out the other. For example, the bottom-up approach could aid the design of top-down models as it allows loans to be classified into pools using individual loan data, whilst the pool-level risk parameter could eventually be estimated from the aggregated data. Since all of the individual card defaults are used to construct the EAD models in this paper, our method would be primarily classified under the bottom-up approaches. Examples of other studies that, similarly to us, utilise a bottom-up method for retail credit card modelling are Tong et al. (2016), Hon and Bellotti (2016) and Leow and Crook (2016).

2.3 Data and variables

The original dataset provides monthly account-level data on the consumer credit cards of a large Hong Kong bank from January 2002 to May 2007. We define EAD as the outstanding balance at default time, taking the amount owed by the borrower excluding any subsequent interests and additional fees; any debt incurred after

default will not be included in the EAD calculation. We say that an account goes into the default state when a borrower either: (1) misses or could not make the minimum repayment amount required by banks for three months or more; (2) is declared bankrupt; or (3) is declared charged-off, i.e. expected to be unable to return the owed money back to the bank. In keeping with the standard practice in EAD modelling, we extract data from the defaulted accounts only, as the estimation is conditional on default and the balances of defaulted and non-defaulted accounts are expected to behave differently.

We also add macroeconomic variables to the dataset because individual customers' borrowing levels could further vary under different economic scenarios. Also, this may help our model be more time-stable and allows us to assess the impact on EAD of downturn scenarios, thus providing a suitable framework for stress testing required by banks applying the A-IRB approach (Kaposty et al., 2017).

We apply the standard yearly cohort method (Moral, 2006) to prepare the data for analysis and set the reference month where the estimation takes place on 1st November of every year. The values of behavioural and macroeconomic covariates are then collected a month prior to the reference month, namely in October, whereas the response, EAD, is recorded at the occurring default time within 12 months following 1st November for each cohort year. In particular, see Figure 2.1, the balance of defaulted accounts at each month from November 2002 to October 2003 is recorded as an EAD value, and then combined with explanatory variables recorded in October 2002 to build a yearly dataset for the period of 2002-2003. This procedure is performed repeatedly for the period of 2003-2004, 2004-2005, 2005-2006, and November 2006 to May 2007. Eventually, we obtain a collection of yearly datasets ready to be aggregated and analysed. Accounts that lack sufficient monthly records to calculate the explanatory variables are omitted.

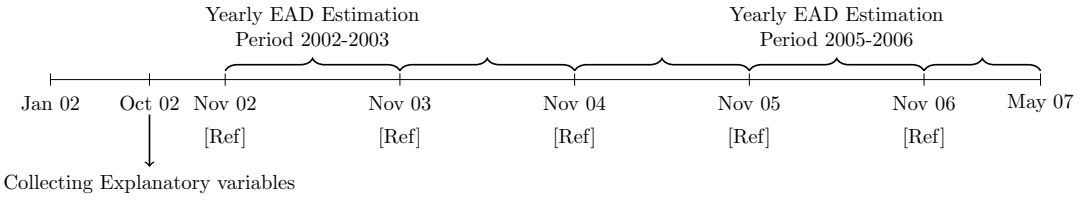


FIGURE 2.1: Standard yearly cohort method applied for EAD dataset.

Further removing a small number of missing value cases (177 observations), we are left with 74,096 defaulted accounts. This dataset is then separated into three groups: training (60%), validation (20%), and test dataset (20%). Figure 2.2 shows the empirical EAD distribution, which exhibits right-skewness, is heavy-tailed, and has a small bump at 200,000 (which is a likely consequence of the bank operating a maximum limit).

An additional analysis of CCF is performed and demonstrated in Figure 2.3. Because of its denominator, the histogram of observed CCF expresses a clear instability; a sheer number of observations are out of [0,1] interval, and the possible range of countable CCF can run freely from the largest value of 97,136 to the smallest value of -88,639 as a result of the contracting denominator. There are 11 accounts with infinite CCF due to its zero denominator. The irregular and unstable shape of observed CCF proves difficulty in modelling task. Otherwise, truncation at zero and one is required in order to comply with the regulation guidance; however, this might follow by a severe information loss. Furthermore, Tong et al. (2016) and Leow and Crook (2016) demonstrated in their work that a direct method of modelling EAD could effectively provide a more accurate predictive performance (e.g. in terms of the mean absolute error) than the conventional CCF approach. Therefore, this paper focuses on a direct EAD modelling without the CCF formulation.

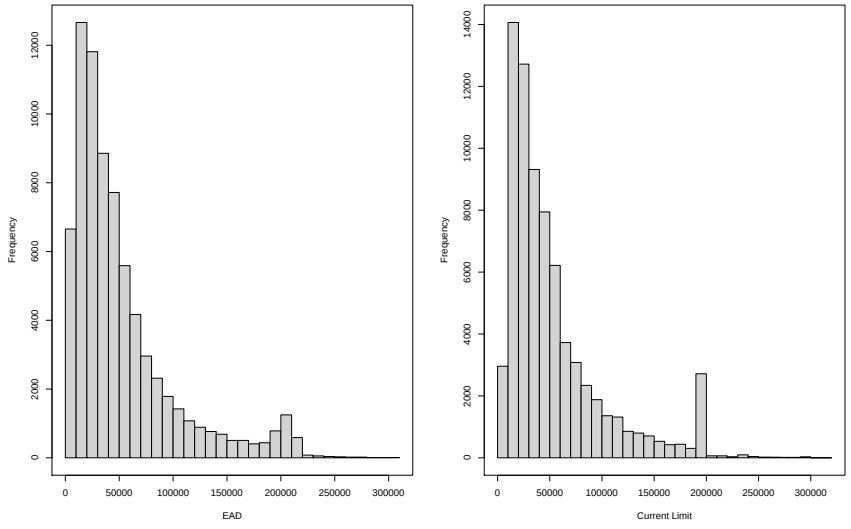


FIGURE 2.2: Histograms of: observed exposure at default (left); observed current limit (right).

Table 2.1 lists the set of candidate explanatory variables extracted from the data, which have previously shown correlation with EAD according to the literature or can be reasonably expected to significantly impact EAD. Four macroeconomic variables are considered: unemployment rate, interest rate, GDP, and CPI.

2.4 Statistical models

The following subsections outline our newly proposed model, GAMLSS.Mix, and three benchmark models, GAMLSS, OLS.Mix and OLS.

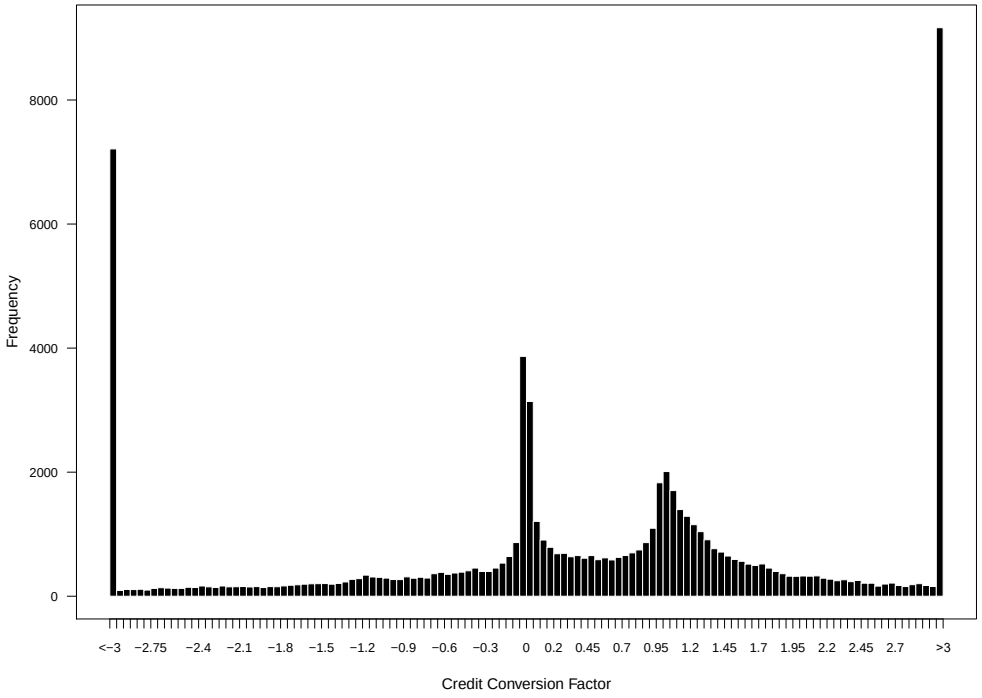


FIGURE 2.3: Histograms of observed Credit Conversion Factor (CCF).

2.4.1 GAMLSS.Mix

In our new model, we propose to estimate EAD conditionally on two mutually exclusive scenarios that may occur in the run-up to default. Denote the EAD of account i as $\text{EAD}_{t,\tau}(i) = \text{EAD}_i$. Note that reference time t and default time τ are omitted from here on for the sake of simplicity. We define a binary variable, S_i, to denote the occurrence of a "max-out" event as:

$$S_i = \begin{cases} 1 & \text{if the balance hit the limit at any point during the outcome window;} \\ 0 & \text{otherwise,} \end{cases}$$

where the outcome window is the period between reference and default time. Applying the law of conditional expectation, the expected value of EAD_i is then given by:

$$E(\text{EAD}_i) \;=\; [P(S_i = 1) \times E(\text{EAD}_i | S_i = 1)] + [P(S_i = 0) \times E(\text{EAD}_i | S_i = 0)]\,. \quad (2.1)$$

Therefore, three model parts must be fitted, all of which conditional on default: first, a model for the probability that the balance will hit the limit over the observation period, $P(S_i = 1)$; second, a model to estimate EAD conditional on the balance hitting

the limit, $E(\text{EAD}_i|S_i = 1)$; third, a model to estimate EAD conditional on no such max-out event occurring, $E(\text{EAD}_i|S_i = 0)$.

We will refer to this newly proposed mixture model as "GAMLSS.Mix", as it will use the GAMLSS framework to fit each of these model parts. For each such component, we use a separate validation set (setting aside 20% of the data) to make model selection decisions such as variable selection. Subsequently, the final model (whose partial residual plots will be shown) is trained after merging training and validation data.

Variable	Notation	Explanation
Age of account	age	Months since account has been opened.
Limit	l	Credit limit, i.e. maximum amount that could be drawn from card.
Balance	b	Current amount drawn.
Behavioural score	bsco	Internal score capturing current credit quality of account.
Months in arrears past 9 months	no.arr9	The number of months an account has been in arrears over the nine months prior to the reference time. A borrower is considered in arrears when they pay less than their monthly minimum payment.
Months in arrears past 3 months	no.arr3	
Limit increase past 9 months	limin9	Dummy variable indicating whether the limit has been increased over the past nine months (Y/N).
Limit increase past 3 months	limin3	
Absolute balance change past 9 months	abs.ch.b9	
Absolute balance change past 3 months	abs.ch.b3	
Average paid percentage past 9 months	paid.per9	Paid percentage is the percentage of last month's balance paid by the borrower, i.e. Paid Amount/Balance.
Average paid percentage past 3 months	paid.per3	
In arrears past 9 months	arr9	Dummy variable indicating whether the account has been in arrears at least once over the past nine months (Y/N).
In arrears past 3 months	arr3	
Credit utilisation	cu	Percentage of the limit drawn by borrower, i.e. Balance/Limit.
Full payment percentage	full.pay.per	Percentage of account's months on book in which borrower has paid balance in full, i.e. number of full payments / age of account.
Behavioural score special code	bscocat	Dummy variable indicating whether behavioural score recorded a "special" case.
(Non-)negative balance	bcat	Dummy variable indicating whether balance was negative (and thus capped).
Time to default	ttd	Duration in months from reference time to default time.
Unemployment rate	unem	HK macroeconomic variable measured at reference time.
Interest Rate	int	The best lending rate benchmarked by the Hong Kong Monetary Authority.
Gross domestic product	gdp	
Consumer price index	cpi	

TABLE 2.1: List of available explanatory variables. Note that, since the behavioural scores of some accounts do not have a regular value (such as 680, 720, etc.) but codes representing "special" cases (e.g., "the account is too new to score"), we replace such special codes by the (training) mean of the regular behavioural scores and flag this up with the help of a dummy indicator (bscocat). Likewise, negative credit card balances, which may e.g. occur when a borrower uses a credit card to purchase a product and decides later to return it, are capped at zero, and another dummy variable (bcat) is added to distinguish between negative and true zero balances.

2.4.1.1 Probability of max-out event

To estimate $P(S_i = 1)$, we model the binary response variable as a non-parametric function of the explanatory variables. More specifically, letting $p_i = P(S_i = 1)$, the max-out event probability is modelled as follows:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_1 Y_{i,t}^T + \alpha_2 Z_t^T + \text{non-parametric terms}, \tag{2.2}$$

where α_1 and α_2 are unknown vectors of parameters to be estimated, and $Y_{i,t}$ and Z_t are (account-level) behavioural and macroeconomic covariate vectors, respectively. The parametric coefficients α_1 and α_2 and non-parametric smoothing terms are fitted by performing the Rigby and Stasinopoulos (RS) algorithm based on penalised (maximum) likelihood (Stasinopoulos et al., 2017), into which the following likelihood function, L, is substituted:

$$L = \prod_{i=1}^{n} p_i^{y_i} \times (1-p_i)^{1-y_i}, \tag{2.3}$$

where $y_i = 1$ for an observation i whose balance hit the limit, and zero otherwise. Penalised B-splines (Eilers and Marx, 1996) are chosen to fit the non-parametric terms in Equation (2.2) because they enable smoothing parameter selection to be performed automatically by minimising the Akaike Information Criterion, AIC = $-2L^p + 2\tilde{n}$, where L^p is the penalised log-likelihood and $\tilde{n}$ is the number of parameters in the model.

We build three candidate models for p_i by considering three different variable selection strategies — either including all explanatory variables or using two alternative stepwise methods (using both forward and backward selection at each step) based on the AIC and BIC criteria. The "gamlss" package (Stasinopoulos et al., 2017) in R (R Core Team, 2020) is used to fit these three models to all training examples of defaults. Based on their performance on the validation set, one of the three candidate models is then selected, following assessments of the Pearson goodness-of-fit statistic from the Hosmer-Lemeshow test (predictive accuracy), Area Under the Receiver Operating Characteristic curve (AUROC) (discrimination power) and residual plots (model adequacy). Where these metrics suggest different candidate models, one is chosen at the modeller's discretion. Note that the residuals used in GAMLSS are normalised quantile residuals which are expected to follow a standard normal distribution regardless of the distribution of the response variable, provided that the model is correctly specified.

Table 2.2 and Figure 2.4 show that while all models demonstrate a good model fit (cf. residual plots), it is the model with full variables that performs best. Hence, we

include all explanatory variables in the probability of max-out event model. All variables, that is not binary, are fitted with non-parametric smoothing terms. (even though some of them express a trend closely to linear, e.g. see the balance effect in Figure 2.8). The binary variables, e.g. bsco and arr9 (see Table 2.1), are modelled with linear terms.

Model	Hosmer-Lemeshow test	AUROC
Full variables	84.68	0.8961
Stepwise AIC	88.62	0.8960
Stepwise BIC	86.22	0.8958

TABLE 2.2: Performance measurements for probability of max-out event model, based on the validation set.

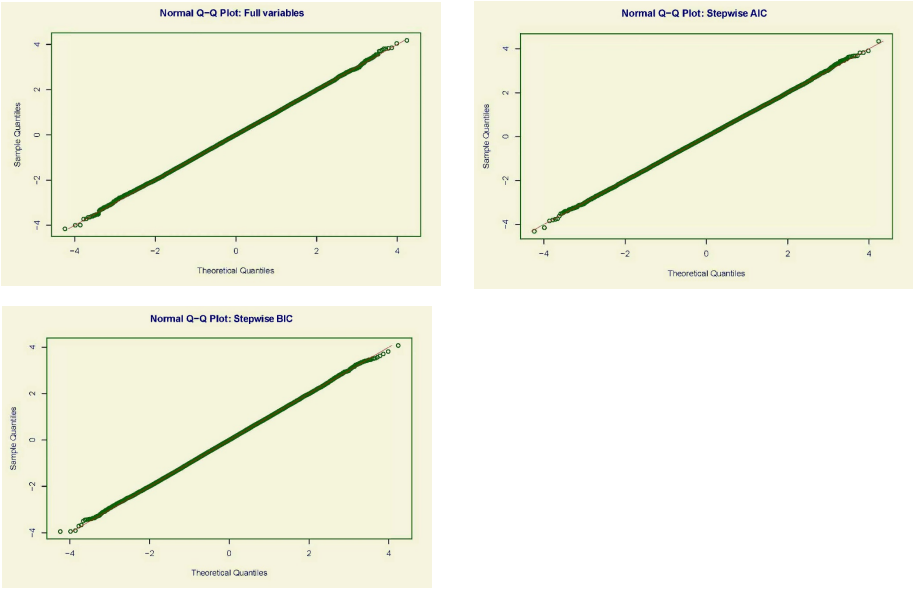


FIGURE 2.4: Residual plots for probability of max-out event model, based on the validation set.

2.4.1.2 Conditional EAD models

To produce EAD estimates that are conditional on either of the two credit balance scenarios, we further partition the training data into two subsets. The first subset consists of the credit card accounts whose balance hit the limit in any of the months during the outcome window; the second subset consists of the accounts that did not. We then proceed by fitting two separate models to these subsets.

In either of these scenarios, one can further distinguish between zero and non-zero EAD values. Zero values may potentially occur because of several special cases or technical default examples, such as charge-offs connected to other accounts, the observations being rounded or truncated to zero, customers moving their outstanding balance to other accounts, or payment delays. There are 427 accounts with zero EAD

in our dataset. As they could have different explanatory drivers, we treat zero values separately from non-zero EAD values by including the probability of zero EAD into the models.

Figure 2.5 shows the empirical distribution of non-zero EADs for both subsets of accounts, confirming that accounts that hit their limit tend to have larger EAD values. Their shape also suggests a positively skewed distribution such as Gamma, Inverse Gaussian, or Log Normal distribution. For each of these candidate distributions, we evaluated the AIC/BIC and MAE/RMSE criteria for a full model (i.e. with all explanatory variables). Based on this, as in Tong et al. (2016), the Gamma distribution was found to give the best results.

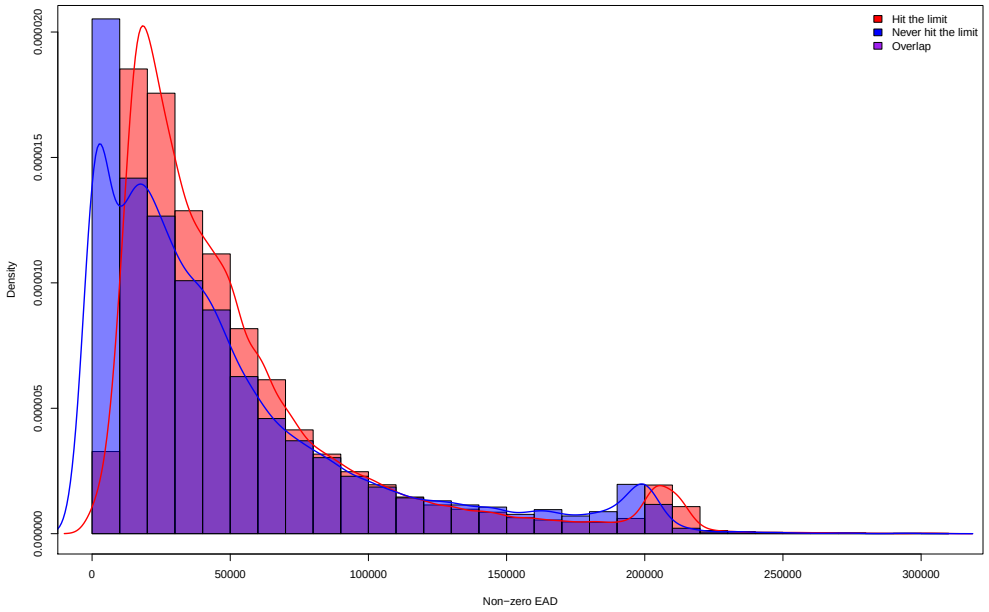


FIGURE 2.5: Empirical distribution of non-zero EADs; red: histogram for the accounts whose balance hit the limit, blue: histogram for the accounts that never hit the limit; purple: overlapping area.

Hence, in order to model $E(\text{EAD}_i|S_i = 0)$, we assume that EAD_i follows a mixed discrete-continuous Zero-Adjusted Gamma (ZAGA) distribution, shown in Equation (2.4).

$$f(\text{EAD}_i|S_i = 0) = \begin{cases} \nu_i & \text{if } (\text{EAD}_i|S_i = 0) = 0, \\ (1-\nu_i)\,\text{Gamma}(\text{EAD}_i|\mu_i, \sigma_i, S_i = 0) & \text{if } (\text{EAD}_i|S_i = 0) > 0, \end{cases} \tag{2.4}$$

for $0 \leq \text{EAD}_i < \infty$, where $0 < \nu_i < 1$, $\mu_i > 0$, $\sigma_i > 0$, and

$$\text{Gamma}(y|\mu,\sigma) = \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{\left(\frac{1}{\sigma^2}-1\right)} e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}.$$

Note that the mean and variance of $\text{Gamma}(y|\mu,\sigma)$ are μ and $\sigma^2\mu^2$, respectively. Hence,

$$\begin{aligned} E(\text{EAD}_i|S_i = 0) &= (1-\nu_i)\mu_i, \\ Var(\text{EAD}_i|S_i = 0) &= (1-\nu_i)\mu_i^2(\sigma_i^2 + \nu_i). \end{aligned} \tag{2.5}$$

There are thus three parameters in the ZAGA distribution: the mean (μ) and dispersion (σ) of non-zero EAD, and the probability of zero EAD (ν). Allowing the relationship between μ and its explanatory variables to be non-linear, we again model it through non-parametric smoothing terms. Since the main focus is on μ, we restrict the relationships of σ and ν with their respective sets of explanatory variables to be parametrically linear. This makes the model less computationally expensive and easier to implement in practice. The parameters μ, σ, and ν can thus be estimated through the following link functions:

$$\log(\mu_i) = \gamma_1^{\mu} Y_{i,t}^{\mu\,T} + \gamma_2^{\mu} Z_t^{\mu\,T} + \text{non-parametric terms};$$

$$\log(\sigma_i) = \gamma_1^{\sigma} Y_{i,t}^{\sigma\,T} + \gamma_2^{\sigma} Z_t^{\sigma\,T}; \quad \text{logit}(\nu_i) = \gamma_1^{\nu} Y_{i,t}^{\nu\,T} + \gamma_2^{\nu} Z_t^{\nu\,T},$$

where γ_1 and γ_2 are unknown vectors of parameters to be estimated. We apply a log and logit link function, respectively, in order to assure that the range of μ and σ parameters are greater than zero and the range of ν parameter is between zero and one. The likelihood function, L, used in the penalised maximum likelihood estimation is:

$$L = \prod_{i=1}^{n} f(\text{EAD}_i) = \prod_{\text{EAD}_i=0} \nu_i \prod_{\text{EAD}_i>0} (1-\nu_i) \times \text{Gamma}(\text{EAD}_i|\mu_i,\sigma_i). \tag{2.6}$$

Five variable selection techniques are applied to create five submodels: using all variables; using stepwise variable selection for μ, σ and ν separately, with either AIC or BIC as the model selection criterion; using stepwise with AIC/BIC by running the parameters together (cf. stepGAICAll.A() function in Stasinopoulos et al. (2017)). The criteria used to select one of the five resulting submodels are Pearson correlation (discrimination performance), MAE, Normalised MAE, RMSE, Normalised RMSE

(predictive accuracy) and residual plots (model adequacy), each of which is again evaluated on the validation set. A normalised version is produced where MAE and RMSE are calculated for EAD/Current Limit, instead of EAD, in order to investigate the performance of the model if the percentage of current limit (not EAD itself) at default time is of interest. Table 2.3 and Figure 2.6 show that while all models demonstrate a similar decent model fit (cf. residual plots), the model with the method of stepwise AIC (run separately) gains the best performance. The chosen stepwise approach suggests excluding the following variables: paid.per9, arr3, arr9, int and cpi (see Table 2.1). The remaining binary variables are fitted with linear terms, whereas those non-binary ones are modelled with non-parametric smoothing terms.

Model	Correlation	RMSE	MAE	Norm.RMSE	Norm.MAE
Full variables	0.8979	22419	12653	0.3249	0.2137
Stepwise AIC (run separately)	0.8985	22414	12649	0.3247	0.2136
Stepwise BIC (run separately)	0.8971	22496	12769	0.3257	0.2146
Stepwise AIC (run together)	0.898	22415	12653	0.3247	0.2137
Stepwise BIC (run together)	0.8937	22861	12879	0.3291	0.2159

TABLE 2.3: Performance measurements for non max-out EAD model, based on the validation set.

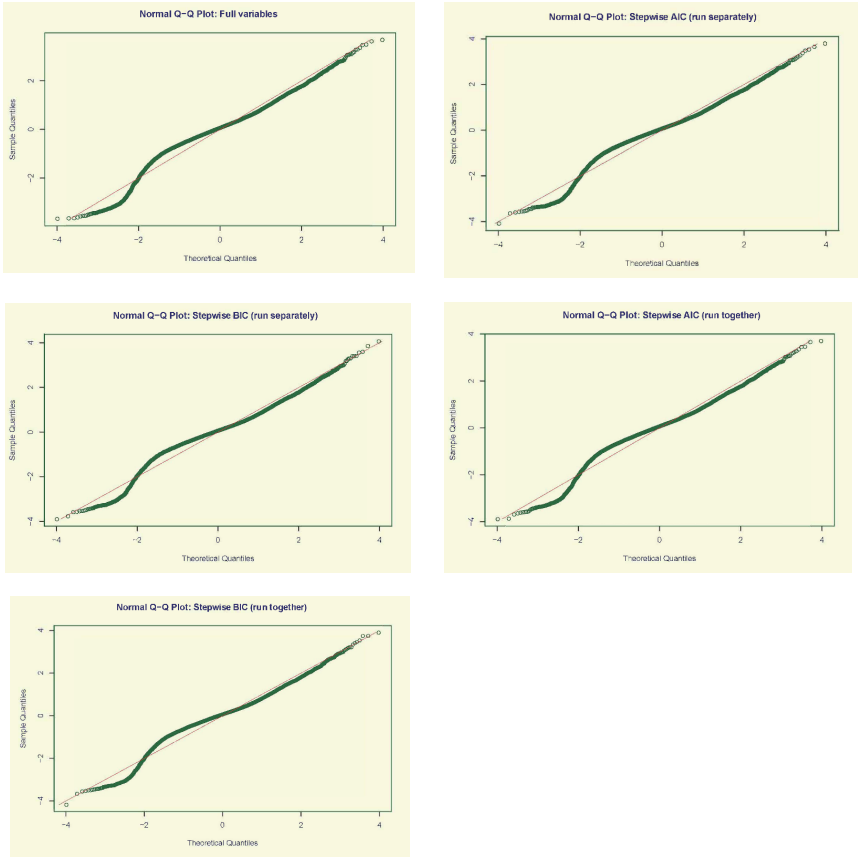


FIGURE 2.6: Residual plots for non max-out EAD model, based on the validation set.

The procedure of modelling EAD for the second subset of accounts that hit their limit, $E(\text{EAD}_i|S_i = 1)$, is similar to its counterpart scenario. The Gamma distribution is again selected for fitting the non-zero EAD response. Note that the best model variations for all three model components can be found in Table 2.4.

Model	Probability max-out	EAD no max-out	EAD max-out
GAMLSS.Mix	Full variables	Stepwise AIC, run separately for each parameter	Stepwise BIC, run separately for each parameter
GAMLSS	Stepwise, with BIC, run for all model parameters together		
OLS.Mix	Stepwise with BIC	Full variables	Stepwise with AIC
OLS	Full variables		

TABLE 2.4: Best submodels for the newly proposed and benchmark models.

2.4.2 Benchmark models

In order to evaluate the effectiveness of our proposed model, we build another three benchmark models against which we compare its predictive performance. Firstly, "GAMLSS" is the EAD model under the GAMLSS framework applied to all defaulted accounts, without applying the mixture idea. Secondly, "OLS.Mix" adds the mixture idea to the OLS framework, applying standard OLS regression for the mixture components and logistic regression for modelling the max-out event probability. Thirdly, "OLS" fits a standard OLS regression model to all defaulted accounts. To perform variable selection for OLS and OLS.Mix, we try three methods: the Least Absolute Shrinkage and Selection Operator (LASSO), a stepwise algorithm, and fitting a model with the full set of variables. As before, we use a validation dataset to find the best (sub)model candidates for each benchmark approach (see Table 2.4).

2.5 Results and discussion

In this section, we present the results of our newly proposed model and the performance comparisons with the benchmark models. In addition, we will inspect the significant relationships between explanatory variables and response parameters.

2.5.1 Discrimination and predictive performance

The performance measurements for all models, evaluated using ten-fold cross validation, are shown in Table 2.5. This table contains the following metrics: Pearson correlation (discrimination performance); MAE, Normalised MAE (see section 2.4.1.2), RMSE, Normalised RMSE (predictive accuracy); and 0.9 quantile loss (QL-90). The α quantile loss function is defined as $\sum_{i;y_i<\hat{y}_i}(\alpha-1)\cdot(y_i-\hat{y}_i)+\sum_{i;y_i\geq\hat{y}_i}\alpha\cdot(y_i-\hat{y}_i)$,

where y_i and $\hat{y}_i$ are true and predicted EAD values, respectively. Its basic idea is to give different penalties to a misestimation based on the selected quantile. The 0.9 quantile loss penalises underestimation more heavily, and hence, is a good measure for assessing the conservativeness of a risk estimate such as EAD.

Model	Correlation	RMSE	MAE	Norm.RMSE	Norm.MAE	QL-90
GAMLSS.Mix	0.937	16927	7881	0.265	0.147	4125
	(0.004)	(385)	(137)	(0.005)	(0.002)	(153)
	0.933	17458	8136	0.273	0.151	4354
	(0.004)	(381)	(140)	(0.006)	(0.002)	(153)
GAMLSS	0.908	20489	8718	0.292	0.160	4457
	(0.041)	(4903)	(268)	(0.010)	(0.002)	(161)
	0.907	20591	8757	0.293	0.161	4471
	(0.041)	(4882)	(268)	(0.010)	(0.002)	(158)
OLS.Mix	0.935	17152	8751	0.304	0.187	4365
	(0.004)	(435)	(161)	(0.011)	(0.003)	(159)
	0.932	17574	8845	0.298	0.183	4532
	(0.004)	(434)	(159)	(0.009)	(0.003)	(154)
OLS	0.930	17810	9758	0.335	0.220	4879
	(0.004)	(397)	(162)	(0.009)	(0.004)	(155)
	0.929	17945	9500	0.315	0.203	4750
	(0.004)	(394)	(147)	(0.007)	(0.003)	(146)

TABLE 2.5: Ten-fold cross validation performance measurements with standard errors inside parentheses; using actual values of time to default (no underline) and weighted approach (with underline).

The variable "time to default" is the number of months from the reference date (1^{st} November) to default date whose range is between one and twelve. It is not the account's age at default time. As time to default is unknown a priori and would not be available in real data for forecasting, Table 2.5 presents two different sets of results: one using the actual values of time to default (to enable comparison with other papers that included this variable and as it is likely to affect dispersion); and one where they were estimated by applying a simplified version of the PD-weighted approach by Witzany (2011), in which, for each month t ($t = 1, \ldots, 12$) of each default cohort, we observe the empirical proportion of training set defaults, $\text{PD}(t_i)$ and, from those, derive the following point prediction for EAD of each account:

$$\text{EAD} = \sum_{t_i=1}^{12} \left[\text{PD}(t_i) \times \text{EAD}(t_i)\right], \tag{2.7}$$

where $\text{EAD}(t_i)$ is the EAD estimate when t_i months is substituted instead of the actual time to default. The latter approach is used to verify to what extent the former performance results remain robust if the model is applied not for explanatory (using real values of time to default) but for prediction purposes (using the estimated values).

Examining the results, we can see that, when it comes to Pearson correlation, there is little to separate the different models, indicating that even the simplest model (OLS) can already discriminate well between high and low EAD risk. However, with regards

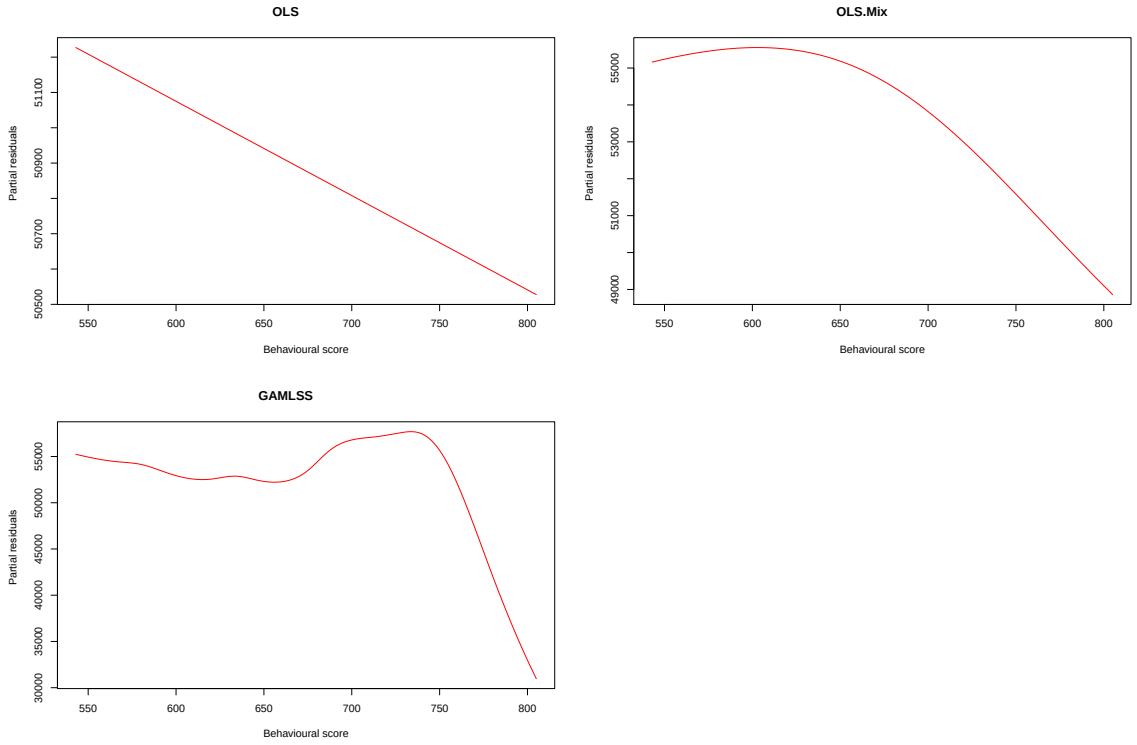


FIGURE 2.7: Partial residual plots of behavioural score vs. estimated EAD, for OLS, OLS.Mix and GAMLSS models.

to all other measures, there are pronounced differences between the various approaches. Firstly, with the exception of RMSE and QL-90 for GAMLSS, the two approaches that apply the GAMLSS framework (GAMLSS and GAMLSS.Mix) outperform those using standard OLS regression (OLS and OLS.Mix), showing that its features are better capable of handling the EAD distribution and its relation to the risk drivers (e.g., any non-linearity). Secondly, when we introduce the mixture concept into the OLS framework (OLS.Mix vs. OLS), all of the predictive accuracy measures improve as well. This is in agreement with the results reported by Leow and Crook (2016), who also found that adding the mixture component to their linear models improved performance. We suggest, as a partial reason for this performance gain, that conditioning on the occurrence of a max-out event has the beneficial effect of introducing some non-linearity into the functional relationships between explanatory variables and EAD. This is illustrated by the partial residual plots for the behavioural score variable in Figure 2.7, showing us how OLS.Mix is able to approximate the non-linear relationship between behavioural score and EAD using a concave function. Thirdly, and perhaps most importantly, the newly proposed model, GAMLSS.Mix, consistently outperforms all benchmark models across all predictive performance criteria (cf. RMSE, MAE, Norm.RMSE, Norm.MAE), whilst being more conservative in terms of the prediction errors it makes (cf. QL-90). This shows that, as hypothesised, there is indeed added value in combining both modelling elements.

When comparing the predictive performance without prior knowledge of time to default (see results with underline) against that of the explanatory model application (i.e. with knowledge of time to default), we see a small drop in performance, as to be expected, but importantly, the performance ranking for all models remains similar and the proposed GAMLSS.Mix model still has the best predictive power. This suggests that our findings are robust regardless of the chosen treatment of this explanatory variable.

2.5.2 Risk Drivers of GAMLSS.Mix model components

Unlike with a linear regression, the non-parametric smooth functions fitted by GAMLSS.Mix cannot be explained in a simple mathematical form; that is, we cannot gauge the impact of an explanatory variable on the response variable by just looking at its estimated coefficient. However, we can display each effect visually with the help of partial residual plots. These depict how one specific explanatory variable influences the response assuming that the other covariates are fixed.

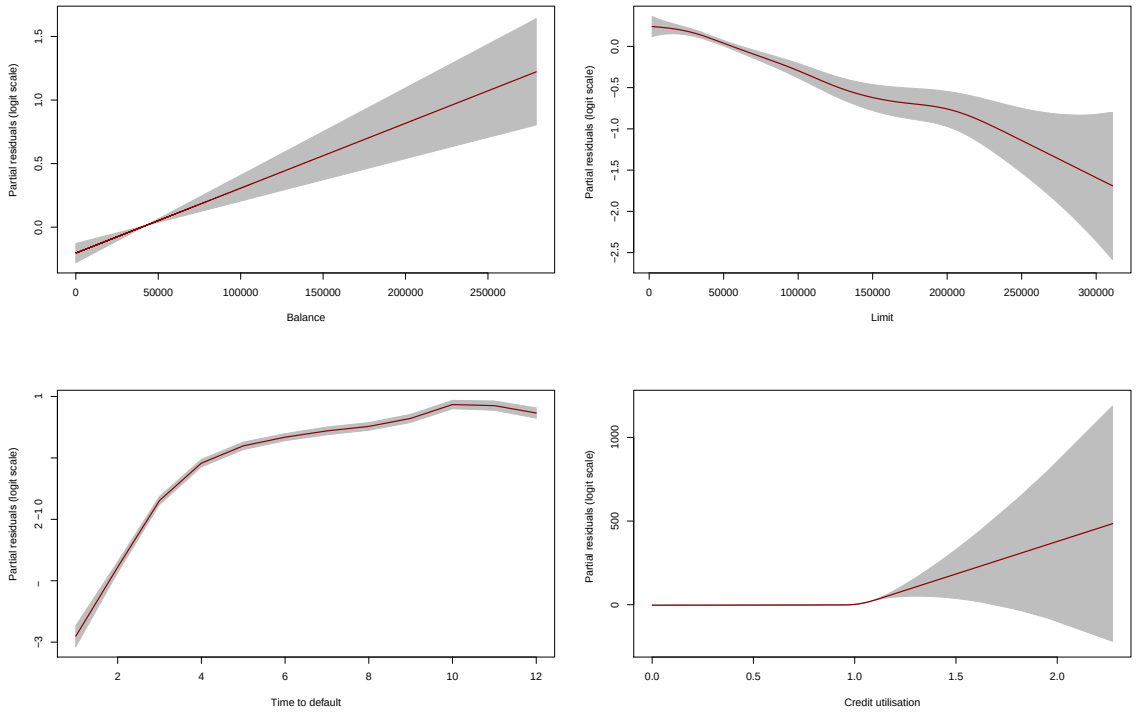


FIGURE 2.8: Partial residual plots on logit scale for max-out event risk in the GAMLSS.Mix model.

Figure 2.8 displays partial residual plots on a logit scale for the max-out event probability, $P(S_i = 1)$, of GAMLSS.Mix. The shaded areas indicate the precision of the estimates using 95% confidence intervals. In the bottom-right panel, we observe that

higher credit utilisation (measured at reference time) makes it more likely that the customer will max out their card in the run-up to default, especially when utilisation already exceeds one prior to the outcome period (the latter makes the event almost inevitable). Similarly in line with expectations, longer time to default (see bottom-left panel) is associated with a higher probability of the balance hitting the limit. Starting balance (top-left) and credit limit (top-right) tend to have a positive and negative effect on the probability of a max-out event, respectively, which is again intuitive since customers with higher balance and lower limit are closer to maxing out their card.

Figure 2.9 presents the partial residual plots, on a log scale, for the μ parameter (non-zero EAD mean) of GAMLSS.Mix, for the subset of accounts whose balance never hit the limit (hence, conditional on $S_i = 0$); Figure 2.10 does so for the other subgroup ($S_i = 1$). In both figures, we see that higher credit limit level is strongly linked to larger EAD. This is again perfectly intuitive as customers with a higher limit are allowed to borrow more. Note that the waviness and widening confidence band near the upper-end of the variable range suggest some undersmoothing linked to the relatively small number of accounts with a limit above 200,000.

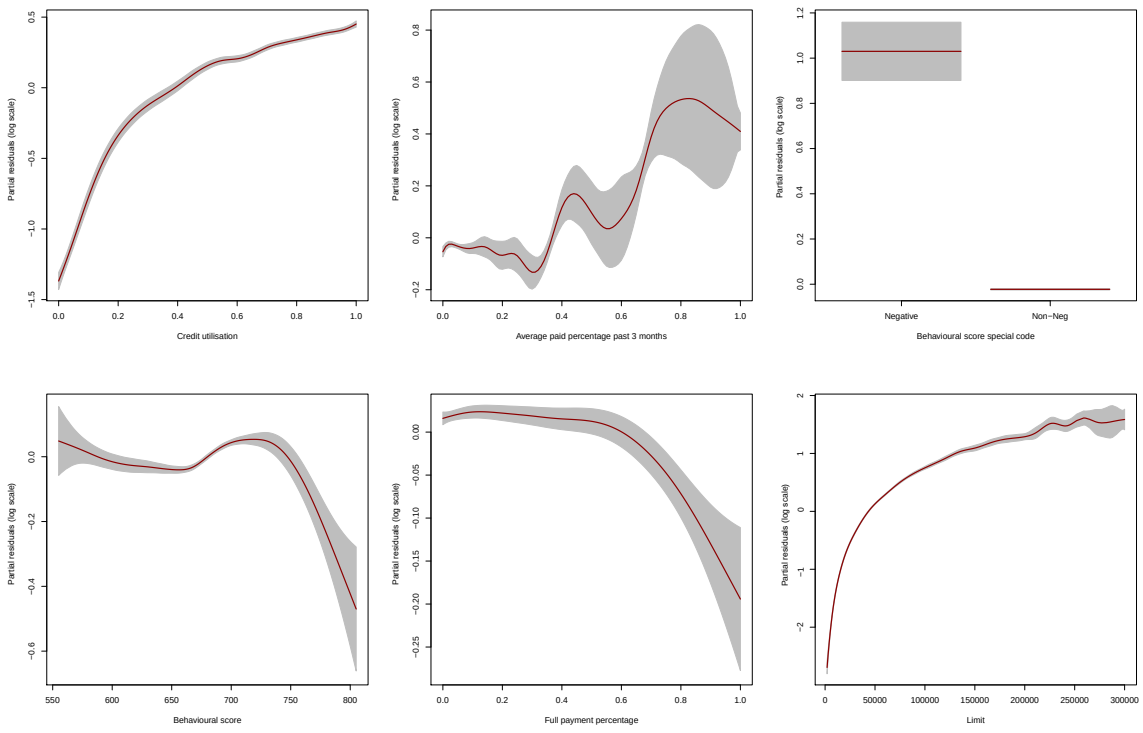


FIGURE 2.9: Partial residual plots on log scale for the mean (μ) parameter of the accounts whose balance never hit the limit in the GAMLSS.Mix model.

Similarly, EAD is also related to the current level of credit utilisation (Figure 2.9, top-left plot) or to balance (Figure 2.10, left plot), higher values implying larger balance at default. Interestingly, more variables appear in Figure 2.9, suggesting that

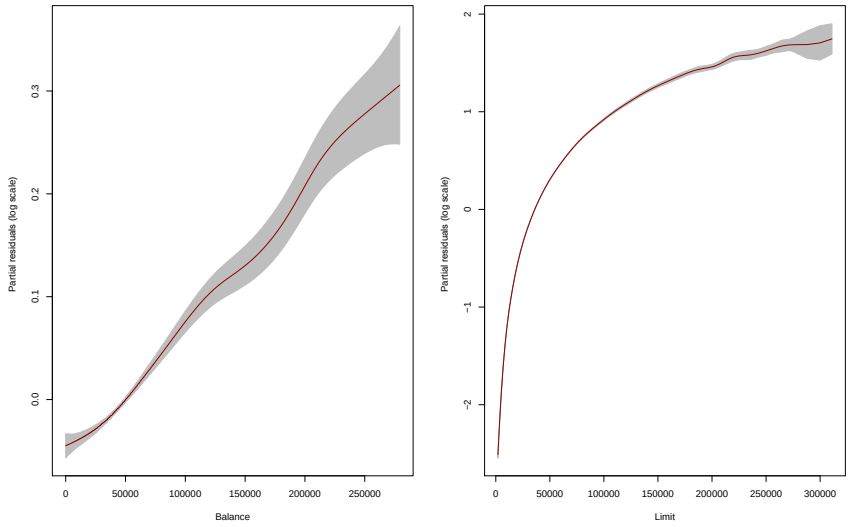


FIGURE 2.10: Partial residual plots on log scale for the mean (μ) parameter of the accounts whose balance hit the limit in the GAMLSS.Mix model.

these only help to better predict accounts who stay clear of the limit. In other words, a more complex model is needed for this mixture component than for the other. For example, in their higher value range, behavioural score and full payment percentage have a negative effect on the EAD of those accounts; hence, provided that they did not hit the limit, high credit-quality borrowers who most of the time pay back their balance in full tend to have a lower balance if they do default. Two novel insights were encountered as well. Firstly, the partial residual plot for average paid percentage over the past three months (see top-middle panel of Figure 2.9) suggests that those borrowers who previously repaid a higher (partial) proportion of their balance could still end up with a higher EAD. Secondly, customers with a negative current card balance (who are thus owed money by the bank) may have higher EAD risk than those with zero balance (top-right). One potential explanation may lie in that both could be seen as indicative of greater card activity. Another may be that, as those values are more often associated with customers who are less likely to default, there may be hidden risks that drive them to heavily draw down before default eventually occurs; this would concur with Barakova and Parthasarathy (2013) who reported that higher EAD can be associated with defaults that are hard to anticipate. Note that, for brevity and as they had a lesser impact (based on a likelihood ratio test), macroeconomic covariates and the other behavioural variables are omitted from the figure (results available on request).

To facilitate further comparison between the different models and the effects they captured, Table 2.6 summarises which explanatory variables are shown to have a strong impact on (non-zero) EAD mean (μ) in the two GAMLSS.Mix component models and the GAMLSS benchmark model and whether that impact is (mostly) positive or negative. Likewise, it also contains the same information for the σ

(dispersion) and ν parameters. For brevity, we omit further discussion of the last parameter, ν.

Variable	Mean			Dispersion			Prob. of zero-EAD		
	EAD	EADn	EADt	EAD	EADn	EADt	EAD	EADn	EADt
age					−			+	
l	+	+	+	+	+			+	
b	+		+	−	−	−	−		
bsco	−	−		−			+	+	
no.arr9					+				
no.arr3								−	
limin9									
limin3									
abs.ch.b9						+	−	−	
abs.ch.b3				+					
paid.per9									
paid.per3	+	+			+			+	+
arr9									
arr3									
cu		+			−			−	
full.pay.per		−						+	
bscocat (special)				+	+				
bcat (negative)	+	+		+	+				
ttd				+	+	+		−	
unem									
int								−	
gdp									
cpi									

TABLE 2.6: A set of strongly significant predictors for the EAD parameters of: the GAMLSS benchmark model (EAD); GAMLSS.Mix no max-out (EADn); and GAMLSS.Mix max-out (EADt).

Turning to the second parameter, dispersion, we can see in Table 2.6 that the higher is the level of credit utilisation and/or current balance, the lower is the dispersion — in other words, the more predictable the EAD. In contrast, the farther away from default time (both scenarios) or the larger the limit (non-max-out scenario only), the larger the dispersion; i.e. there is more time and scope for the balance to change and thus become less predictable. These four effects all appear to be intuitive. Interestingly, as for the EAD mean earlier, the list of important factors is again longer for the first mixture component (i.e. for the accounts with no recorded max-out event). There, age of account (i.e. time on book), the average of paid percentage over three months, and number of months in arrears are also among the variables that are shown to affect dispersion. Specifically, the longer the account has been on the books, the more predictable is EAD, whereas higher values for the other two variables (which could indicate greater monthly variation in balance) tend to imply greater variance. Also, special behavioural scores and negative current balances imply special cases under

which the EAD prediction for those accounts becomes more uncertain as well. As there are all meaningful effects, there appears to be added value in explicitly modelling the dispersion parameter (rather than assuming homoscedasticity).

2.6 Conclusions and future research

Exposure at Default (EAD) is one of the key parameters used to calculate the regulatory capital requirements under the Advanced Internal Rating Based (A-IRB) approach. To estimate EAD, Credit Conversion Factor (CCF) models were implicitly suggested by the Basel Accords and have been studied in the literature, but several drawbacks of such models can prove problematic. In this paper, we therefore mainly focus on estimating EAD via a direct model rather than applying CCF or other related factors.

Our newly proposed model combines two ideas formerly put forward in the literature. First, it is built under the GAMLSS framework which produces a much more flexible fitted distribution than the GLM and GAM frameworks. Second, as the level of EAD as well as the risk drivers of its mean and dispersion parameters could significantly differ depending on whether the account hit the credit limit at any point in the run-up to default, we extend our solution to a mixture model conditioning on these two possible scenarios. This new model, as well as several benchmark models, are empirically validated using a large dataset of credit card defaults not previously used in the EAD literature.

By distinguishing between these two scenarios, we indeed found differences in preferred risk drivers for the EAD model parameters. For example, current balance was picked over several other potential drivers for (positive) EAD mean when a max-out event occurs, but not in the opposite scenario, whereas current limit level was identified as being strongly linked to dispersion only under the non-max-out scenario. Moreover, the number of factors is larger for borrowers who did not max out their cards, suggesting that this subgroup benefits from a more complex model. Overall, only behavioural variables appear to have a significant impact in our EAD models; despite the data containing defaults from a recessionary period, the macroeconomic covariates show little added predictive power over those account-level variables. Current limit is the strongest variable that affects the mean of non-zero EAD. To manage model uncertainty, one should focus on the current level of drawn balance amount and (estimated) time to default as their values greatly impact EAD dispersion.

Our results show a clear performance benefit of applying GAMLSS over the OLS framework, confirming, consistently with what Tong et al. (2016) reported for another dataset, that there are indeed predictive accuracy gains in EAD modelling from including non-linear effects and targeting not only the EAD mean but also dispersion.

Similarly, when the mixture concept is introduced into the OLS framework, all predictive accuracy measures improve as well. A new explanation we put forward for the latter is that, by implementing the mixture idea, we allow some non-linear effects to emerge from the combination of different linear models, thus capturing more complex relationships between EAD and its covariates and producing better predictions. Most of all though, we find that combining the mixture component and the GAMLSS approach results in another predictive performance boost, as our newly proposed model, GAMLSS.Mix, outperforms the three benchmark models on all criteria.

In terms of potential practical benefits, a more accurate EAD model, such as that proposed, can lead to more accurate loss estimation, which allows banks to adjust the capital they require accordingly. Moreover, the non-linear predictor effects, shown in the partial residual plots, reveal the impact of each behavioural variable on different risk aspects. This can provide the bank with useful insights as it designs an early warning system. More specifically, the insights from the "max-out" model allow the bank to identify those borrowers who are most at risk of maxing out their credit card (and thus present the largest exposure risk). It follows that the bank could decide to lower their credit limit to mitigate such risk.

A potential future avenue of research is to more fully incorporate time to default in the prediction framework, particularly since our models confirm that max-out risk and EAD variance (dispersion) are higher the more time elapses before default. As time to default is unknown a priori, one could use survival analysis to capture its dynamic distribution, from which EAD can then be derived as in section 2.5.1. A follow-up study could consider different methods to implement such a PD-weighted approach (Witzany, 2011) and test their effectiveness when combined with the newly proposed EAD model.

Chapter 3

An Additive Copula Regression Model for Credit Card Balance and Probability of Default

Suttisak Wattanawongwan [a,1], Christophe Mues [b], Ramin Okhrati [c], Taufiq Choudhry [b], Mee Chi So [b]

[a] School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[b] Southampton Business School, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[c] Institute of Finance and Technology, University College London, London, WC1E 6BT, UK

[1] Corresponding author
Email address : S.Wattanawongwan@soton.ac.uk

Abstract

Previous studies have shown that the Basel regulatory capital requirement can be underestimated by ignoring the dependencies between the Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD). In retail credit risk, only a small number of papers have directly modelled account-level dependence between PD and LGD, but no such work has been done yet for the relationship between PD and EAD. To close this gap, we propose a joint model for PD and EAD, evaluating a variety of copulas under the bivariate Copula Generalised Additive

Models for Location, Scale and Shape framework. Using a large dataset of credit card accounts, we explicitly model card balance of both defaulted and non-defaulted accounts, rather than balance at default only. In addition to identifying the dependence structure between default risk and future balance, and the key drivers in each model component, the analysis shows that our proposed model produces a more precise and conservative expected loss estimate compared to other models.

3.1 Introduction

Three parameters are of particular interest when quantifying the credit risk linked to consumer lending: Probability of Default (PD), i.e. the risk that a borrower will no longer be able to satisfy their repayment obligations; Exposure At Default (EAD), i.e. the amount owed by the borrower when they default; and Loss Given Default (LGD), i.e. the percentage of this amount that the lender will be unable to recover. These parameters are used for the calculation of the risk-sensitive regulatory capital requirement introduced by the Basel Accords. Basel's Advanced Internal Ratings-Based (A-IRB) approach employs an asymptotic single risk factor (ASRF) model, as pioneered by Vasicek (2002), to translate the three parameters into the amount of capital required for credit risk. However, this model overlooks any dependence between PD, LGD and EAD, contrasting with a growing number of empirical studies showing a positive relation between them. This can lead to the underestimation of portfolio credit loss and thus capital shortfalls; hence, modelling the interrelationship between these three risk parameters is essential.

In the retail credit risk modelling literature, the dependence between PD and LGD has received little attention, and even less so the relationship with EAD. The latter is of particular interest for credit cards (or other forms of revolving credit), as they allow borrowers to draw more money in the run-up to default. We aim to close such gap by modelling the joint distribution of credit card PD and EAD, considering their dependence under the bivariate Copula Generalised Additive Models for Location, Scale and Shape (CGAMLSS) framework (Marra and Radice, 2017a). This combined flexible framework allows response variables to assume any of a wide range of parametric distributions and their relationship to follow one of various dependence structures selected via a copula function. Also, whereas most previous work has built EAD models on just the subset of defaulted accounts, our approach explicitly addresses potential sample selection bias by extending the analysis to outstanding balance over a 12-month period in a sample of both defaulted and non-defaulted accounts.

Our models are fitted to a large dataset of credit card accounts from a Hong Kong lender. In our analysis, we will identify the key drivers of default risk, balance and

their dependence. To empirically validate the effectiveness of introducing the dependence, we also construct two standalone models, for PD and balance separately, against which we benchmark our newly proposed copula model. Furthermore, we will show how the proposed approach leads to better expected loss estimates at the portfolio level.

The paper is structured as follows. Section 3.2 reviews the relevant literature. Section 3.3 explains the data and variables used, and Section 3.4 describes how the statistical models are constructed. The results are analysed in Section 3.5. Section 3.6 concludes.

3.2 Literature review

In our discussion of the literature, we focus on four streams of work: previous work considering the dependence between PD, LGD and EAD; copulas and their applications to other related settings; existing EAD modelling approaches for credit cards and potential sample selection bias resulting from them; the modelling framework used in this paper (CGAMLSS). At the end, we will summarise how our work advances this body of literature.

3.2.1 Dependencies between credit risk parameters

A growing number of studies have identified dependencies between PD, LGD and EAD. Firstly, it has been found that PD and LGD tend to move in the same direction; a higher default rate is more likely associated with a higher loss rate (Altman et al., 2005; Caselli et al., 2008; Chava et al., 2011; Jacobs and Karagozoglu, 2011; Pykhtin, 2003). As a partial explanation, Allen and Saunders (2003), Frye (2000) and Hillebrand (2006) noted that default and recovery rates are driven by the same macroeconomic variables, as the value of collateral assets (which affects both parameters) depends on the state of the economy. Secondly, a positive correlation between default risk and EAD has been reported by Agarwal et al. (2006), Jiménez et al. (2009), Mester et al. (2006) and Norden and Weber (2010). They suggested that borrowers who are facing financial distress and later default tend to also draw more money from credit cards or lines of credit than those who do not default. Inversely, Araten and Jacobs (2001) and Jacobs (2010) found that lenders would often reduce the credit limit of corporate credit lines if they foresee a pending default, which implicitly implies a reduction in EAD risk for such portfolios.

By neglecting these risk dependencies, portfolio risk can be underestimated since they are the underlying force that significantly increases tail losses. For example, Barco (2007), Miu and Ozdemir (2006) and Rösch and Scheule (2008) ran a series of factor

models and analysed by how much the LGD must be scaled up in order to compensate for the absence of PD and LGD dependence in the ASRF capital formula. They found a large mark-up is required to avoid dramatically underestimating the capital requirement. Bade et al. (2010) and Rösch and Scheule (2014) similarly showed that portfolio credit risk is severely underestimated if the PD and LGD are assumed independent. By performing a series of simulations on factor models, assuming a predetermined set of systematic factors, Kaposty et al. (2017) and Kupiec (2008) found increased tail risk at the portfolio level when EAD is treated as dependent as well. Using downturn estimates for LGD and EAD, proposed by the Accords as the means to alleviate this problem (see pages 96 and 97 in BCBS (2017)), is not a satisfactory solution according to Kaposty et al. (2017), as they found that this can still lead to capital underestimation compared to the model in which PD, LGD and EAD are stochastically dependent on each other.

3.2.2 Copulas and their applications to financial risk

As the preceding discussion shows, accounting for the dependence between the different risk parameters is crucial to avoid underestimating portfolio risk; this suggests a role for directly modelling such dependence at the account level. In the literature (see e.g. Klein et al. (2015)), modelling two or more correlated outcomes given a set of predictors usually relies on a particular multivariate distribution assumption, such as a bivariate Gaussian. The latter implies that both dependent variables must follow a normal distribution with a symmetric dependence structure. However, in the credit risk setting, the response variables of interest may not be Gaussian or follow the same distribution. For instance, EAD always exhibits positive skewness, whereas LGD is bimodal and peaked at zero and one. Also, the correlation between two risk parameters might be stronger at their higher levels, implying an asymmetric structure.

A more flexible option for such scenarios is to construct a joint distribution via a copula-based model. Copula functions enable a multivariate response to be jointly constructed from parametric marginal distributions (Sklar, 1959) which are not restricted to the standard Gaussian or exponential families. Moreover, they allow various dependence structures for the responses via different copula specifications. A key attraction of copulas is that the functional forms of a copula and its components (univariate marginal CDFs) can be specified separately. It follows that one can always construct the joint distribution from arbitrary marginal CDFs by implementing an appropriate copula function. This contrasts to a conventional parametric specification where a joint and marginal distributions need to be known a priori.

In finance, copulas have gained increasing popularity over the past decades (Embrechts et al., 2003; Nelsen, 2006). In the insurance setting, Krämer et al. (2013)

calculated the expected policy loss by jointly modelling insurance claims frequency and claim size via a copula function. They concluded that ignoring their dependence could lead to substantial bias in total loss estimation. Moreira (2010) proposed an alternative approach of calculating the capital requirement, using Clayton copulas to reflect the right tail dependence of PDs; this better explained the extreme loss in adverse scenarios. Calabrese et al. (2019) considered the dependence between defaults in peer-to-peer lending and those observed by credit bureaus, using copula methods and generalised extreme value regression. They found that by connecting these two correlated default risks, the calibration performance of the predicted peer-to-peer loan PD is enhanced. Bade et al. (2010) studied the dependence between default time and LGD, assuming that they are linearly correlated. Krüger et al. (2018) further analysed the dependence between (multi-year) time to default and LGD by means of copulas and found that the lifetime expected credit loss under IFRS 9 increases when this dependence is considered.

Despite ample literature showing the importance of capturing dependence between PD, LGD and EAD, few of these papers focus on building empirical models for these dependencies, particularly those including loan- or account-level covariates. Fewer still have considered a realistic asymmetric dependence structure, apart from Krüger et al. (2018) who modelled PD-LGD dependence using copula methods. To our knowledge, there is no such work yet for the relationship between PD and EAD. To close this gap, we will model PD, EAD and their dependence structure, in the context of a credit card portfolio.

3.2.3 Existing EAD models for credit cards and the problem of sample selection bias

Credit cards have received limited attention in the credit risk literature. Much of the work on EAD has thus far focused on corporate credit, whilst fewer studies address retail customers (Gürtler et al., 2018). This is notwithstanding that credit cards make up the largest share of revolving credit for most A-IRB banks and contribute the largest number of defaults (Qi, 2009).

Previously, Tong et al. (2016), Leow and Crook (2016) and Qi (2009) modelled EAD for consumer credit cards, applying a range of different methods. However, they did not explicitly model the dependence between PD and EAD, nor did they explicitly address potential sample selection bias, as the data they used was derived only from defaulted accounts. This is a common practice to ensure that the predicted balance is indeed conditional on default, but it also means that a much larger volume of non-defaulted accounts, which are necessary to estimate PDs, are neglected. This leads to a potential sample selection bias problem, since the model will be applied to produce loss estimates for the entire portfolio. Addressing the similar problem but for

LGD, using Moody's Default and Recovery Database, Krüger et al. (2018) showed that by modelling LGD based only on defaulted accounts, the capital requirements can be significantly underestimated.

Rather than focusing solely on balance at default time (or equivalently, EAD), we therefore extend the analysis to outstanding balance over a 12-month period in a sample of both defaulted and non-defaulted credit card accounts. Predicting balance for either group is beneficial for expected profit estimation as well as risk management, as argued by Hon and Bellotti (2016), who, similarly to us, modelled the balance of all accounts. However, they did not explicitly identify how to meet the Basel's requirement of having to provide an estimate of balance conditional on default, or how to deal with default time not being known at the time of estimation. In contrast, our approach exploits the joint distribution between default condition and balance and produces not only balance but also the conditional expectation of balance given default (and hence, EAD).

3.2.4 Bivariate Copula Generalised Additive Models for Location, Scale and Shape

We fit marginal distributions for default risk (PD) and credit card balance under the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017), which was previously applied to EAD modelling by Tong et al. (2016). This framework allows a response variable to assume a wide range of parametric distributions, allowing their parameters (location, scale, and shape) to be modelled as a function of predictors using additive terms. This flexibility means we can select a distribution outside of the exponential family to model balance, which has been shown to be right-skewed (Hon and Bellotti, 2016), as well as include any non-linear variable effects into the models. Then, we bind the respective marginal GAMLSS distributions for PD and balance via copulas, to produce a joint bivariate response with a suitable (Gaussian or non-Gaussian) dependence structure. As a result, conditional distributions can be derived from the joint distribution allowing us to see how the two responses affect each other. The dependence parameter(s) of a copula function can as well be modelled as a function of explanatory variables, allowing one to also predict and explain the dependence between the two response variables.

All model parameters (i.e. marginal and copula dependence parameters) are estimated simultaneously, using a recent computational method for fitting bivariate Copula models for the GAMLSS class (referred to from here on as the CGAMLSS framework). Under this extended framework, response margins and copulas are independently and flexibly selected. The coefficient estimation is achieved by maximising a penalised likelihood function and applying the trust region algorithm

(Marra and Radice, 2017a) along with the technique of automatically selecting smoothing parameters. In our model, the copulas explain the joint movement between PD and balance after controlling for covariates' effects on marginal models. For example, current balance may determine the level of both default probability and future balance and thus affects their observed dependencies. Any remaining stochastic effects that cannot be captured by such observable covariates are captured by the copulas. To the best of our knowledge, this paper is the first to apply the CGAMLSS framework in the credit risk area.

3.2.5 Research contributions

To summarise, the main contributions of the paper are that it: (1) shows how to avoid potential sample selection bias in EAD modelling by incorporating both defaulted and non-defaulted credit card accounts; (2) proposes and tests a flexible copula regression approach, new to credit risk, to simultaneously model PD, card balance and their dependence structure; (3) gives further insights into the drivers for each and any remaining dependence between them; (4) demonstrates how this novel approach produces more accurate and conservative expected loss estimates on a real-world credit card portfolio.

3.3 Data and variables

The original dataset provides monthly account-level data relating to the consumer credit cards of a large Hong Kong bank from January 2002 to May 2007. We specify that an account goes into the default state when a borrower either: (1) misses or could not make the minimum repayment amount for 90 consecutive days or more; (2) is declared bankrupt; or (3) is declared charged-off, i.e. expected to be unable to return the owed money back to the bank. The second response variable, balance, is the drawn amount on the card measured at the observation point.

In keeping with the standard practice in EAD modelling, we apply the yearly cohort method (Moral, 2006) to prepare the data, setting the reference month where the estimation takes place to 1st November of every year. All explanatory variables are calculated for the time period leading up to that point. Accounts that lack sufficient monthly records are omitted. For each yearly cohort, a binary variable then indicates whether a default event has been recorded on the account, over the cohort's 12-month outcome window. The value of balance is captured differently depending on this default status. For an account that defaults within the 12-month period, it is measured as the outstanding balance at default time, whereas for non-defaulted accounts, we select the value from the series of observed monthly balances. To avoid potential bias

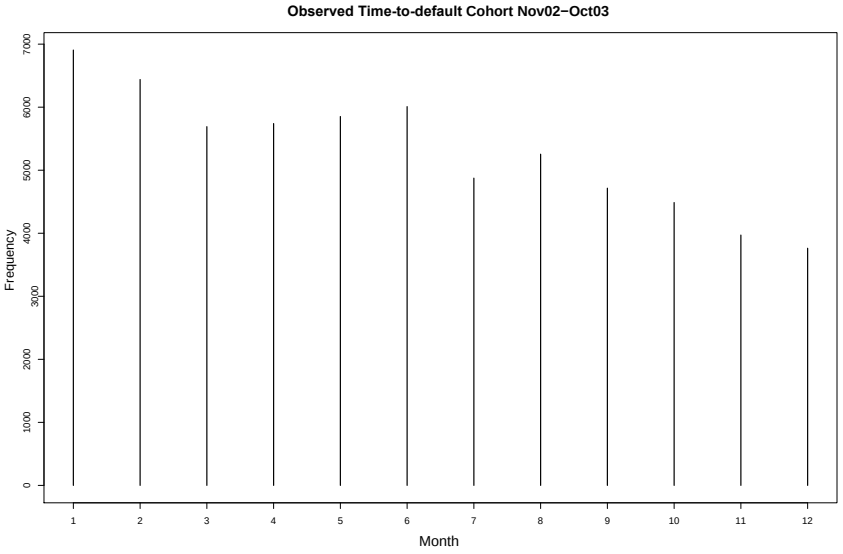


FIGURE 3.1: Yearly empirical time-to-default distribution: November 2002 - October 2003 cohort example.

in picking this observation time and ensure that observations from both subpopulations are maximally comparable, we randomly select a month according to the empirical distribution of time-to-default (i.e. duration in months from reference time to default time) observed for defaulted accounts (see Figure 3.1, for an example cohort period). This method is preferred over taking the balance at the end of the 12-month period, as the latter would risk artificially inflating the balance value for non-defaulted accounts.

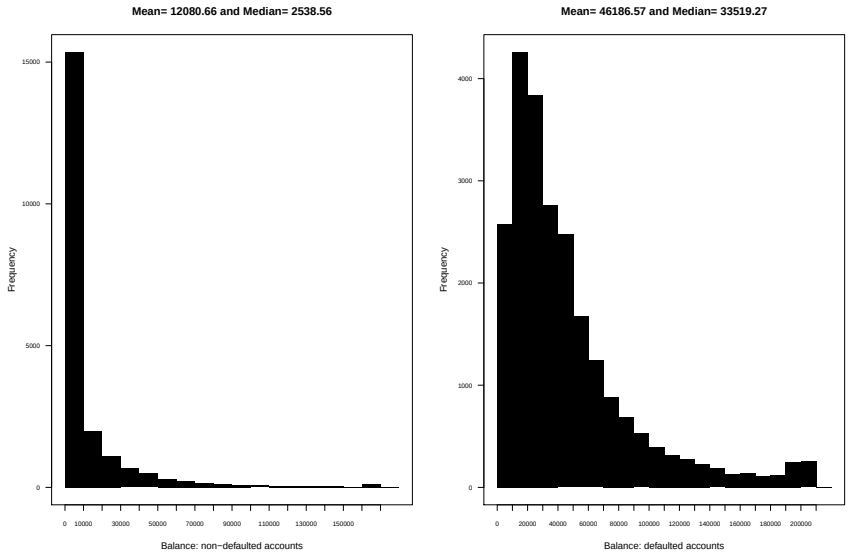
Our dataset contains a low number of defaults (1%) which is not uncommon in practice, especially for banks that exhibit an overall good quality of borrowers (Pluto and Tasche, 2011). Low default portfolios can be problematic in several ways. For instance, a few number of defaults could impact on the model's ability to correctly discriminate between the two outcomes. In other words, the resulting classification model could give different default ratio from the real-world applications (Saerens et al., 2002). In addition, PD estimates could be underestimated and hence do not involve a sufficient conservatism imposed by the regulation (Pluto and Tasche, 2011). We therefore take a balanced sample of the data, reducing the proportion of defaults from 1:99 to 50:50. The balanced sample is produced by first randomly selecting 50% of the available defaulted accounts. Let assume that we have randomly picked x defaults. Then, we randomly select the non-defaulted accounts for the same amount of x. By not using all defaulted data, this procedure is to ensure that both defaults and non-defaults would be selected without bias. In summary, there are 29,303 accounts for both defaulted and non-defaulted accounts used in this study. To rescale the estimated PDs, one can use the Bayes' theorem, as described e.g. in Saerens et al. (2002). Further study could continue to investigate how the choice of sampling affects the model quality and the joint estimates of PD, EAD and outstanding balance.

Variable	Notation	Explanation
Age of account	age	Months since account has been opened.
Limit	l	Credit limit, i.e. maximum amount that can be drawn from card.
Balance	b	Current amount drawn.
Behavioural score	bsco	Internal score capturing current credit quality of account.
Average paid percentage past 9 months	paid.per9	Paid percentage is the percentage of last month's balance paid by the borrower, i.e. paid amount/balance.
In arrears past 9 months	arr9	Dummy variable indicating whether the account has been in arrears at least once over the past nine months (Y/N).
Credit utilisation	cu	Percentage of the limit drawn by borrower, i.e. balance/limit.
Full payment percentage	full.pay.per	Percentage of account's months on book in which borrower has paid balance in full, i.e. number of full payments / age of account.

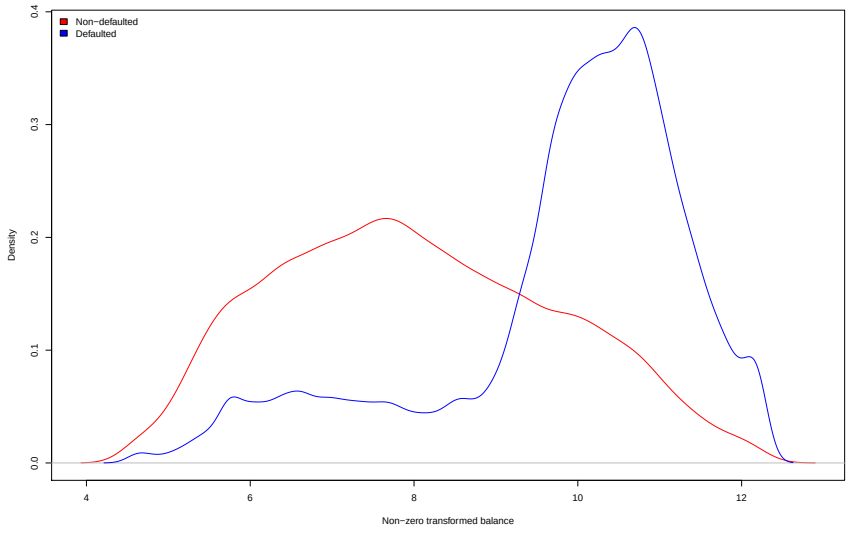
TABLE 3.1: List of available explanatory variables.

Table 3.1 lists a set of explanatory variables that were shown to have a significant relationship with default status, balance, and/or EAD according to previous literature (see e.g. Tong et al. (2016)) and preliminary analysis. Further removing a small number of missing value cases, we are left with 58,606 accounts. This dataset is then separated into three groups: training (60%), validation (20%) and test set (20%).

Figure 3.2a shows the empirical distributions of balance for defaulted and non-defaulted account observations, both of which are right-skewed and heavy-tailed. The higher median balance for the defaults suggests that, similarly to Jiménez et al. (2009) and Mester et al. (2006), there is a strong dependence between default status and the level of balance. A substantial proportion of account observations have zero balance (representing 30% of the dataset), whilst some have extremely large values. This distribution shape is difficult to model, which could result in poor predictive performance. We address this problem by splitting zeroes (including negative values, which were capped at zero) and other values into two separate groups and modelling the probability of balance being zero. To the remaining non-zero values, we then apply the log transformation. As seen from Figure 3.2b, the non-zero log-transformed balance remains larger on average for the defaulted accounts, but with a less pronounced tail. The empirical rank correlation between binary default outcome and balance (Kendall's tau of 0.49) also shows a positive correlation between the two responses.



(a) Histograms of observed balance (one per cohort period) for non-defaulted (left) and defaulted accounts (right).



(b) Density plots of observed non-zero log-transformed balance for non-defaulted (red) and defaulted (blue) accounts.

FIGURE 3.2: Empirical distributions of balance for defaulted and non-defaulted account observations.

3.4 Statistical models

In this section, the model specifications for the standalone and copula models are presented. First, we fit a PD model, and a second model to predict (non-zero, transformed) balance, each under the GAMLSS framework. Their respective coefficients are thus estimated independently of each other. Then, the best standalone PD and balance model specifications, evaluated from the validation data, will be used

as the margins of a bivariate copula model under the CGAMLSS framework, the coefficients of which are estimated simultaneously.

3.4.1 Marginal specification: PD model

To estimate the probability of default, we define $\pi = P[Y_1 = 1]$, where Y_1 denotes default status, taking the value of one when accounts default, or zero otherwise. Logit, Probit and Complementary log-log (Cloglog) are three candidate models for the binary response Y_1. Although the other link functions perform fairly similarly (see Table 3.2), based on its performance on the validation set, the Logit specification is chosen, following assessment of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), Brier score (predictive accuracy) and Area Under the Receiver Operating Characteristic curve (AUROC) (predictive discrimination). Note that we report the mean Brier score and AUROC level computed across ten different subgroups of the validation set, to ensure the results are reliable and not overly influenced by a single data point. The residual plots in Figure 3.3 show a good model fit for all specification options.

Link Function	AIC	BIC	Brier score	AUROC
Logit	23000.24	23368.50	0.1049	0.9224
Probit	23031.39	23369.73	0.1052	0.9221
Cloglog	23152.95	23538.44	0.1058	0.9217

TABLE 3.2: Performance measurements of the candidate marginal distributions for default status Y_1, assessed on the validation dataset.

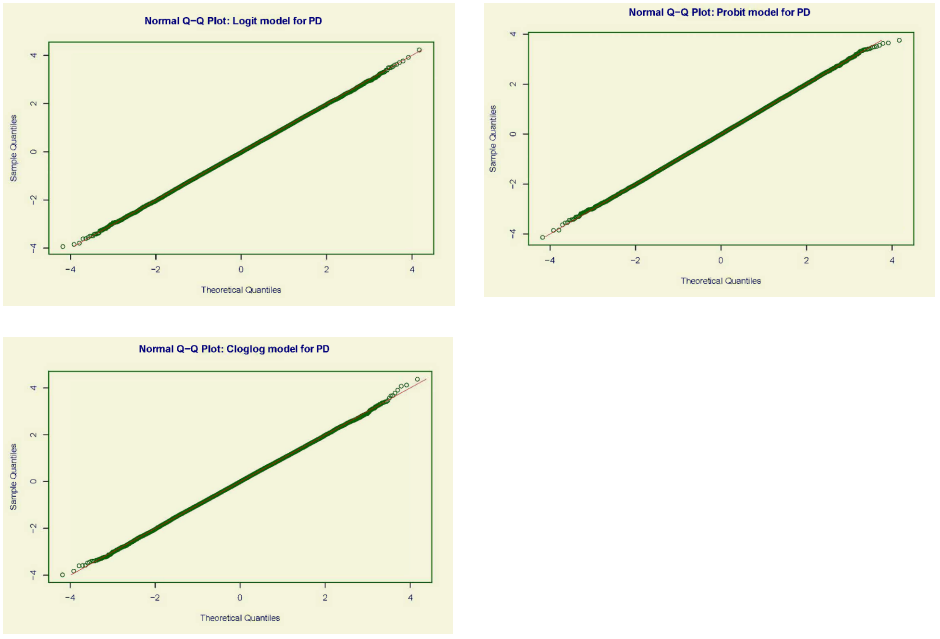


FIGURE 3.3: Residual plots for PD model.

Hence, the default probability, π, where $0 < \pi < 1$, is modelled as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \eta_\pi = \beta_0^\pi + \beta_1^\pi \text{age} + \beta_2^\pi \text{arr9} + s_3^\pi(\text{l}) + s_4^\pi(\text{b}) + s_5^\pi(\text{bsco}) + s_6^\pi(\text{paid.per9}) + s_7^\pi(\text{cu}) + s_8^\pi(\text{full.pay.per}), \tag{3.1}$$

where $\boldsymbol{\beta}$ are parametric coefficients and $\mathbf{s}(\cdot)$ are non-parametric smoothing terms. Existing literature and the authors' previous research (see Tong et al. (2016)) suggest that the variable age (i.e. account tenure) can be modelled with a linear effect, while the other continuous variables are expected to assume a non-linear relationship with logit π. The coefficients $\boldsymbol{\beta}$ and $\mathbf{s}(\cdot)$ are fitted by performing the Rigby and Stasinopoulos (RS) algorithm (Stasinopoulos et al., 2017) and implemented in the R package gamlss (Stasinopoulos et al., 2019), based on penalised (maximum) log likelihood, L^p, into which the following likelihood function, L, is substituted:

$$L = \prod_{i=1}^{n} \pi_i^{y_{1,i}} \times (1-\pi_i)^{1-y_{1,i}},$$

where $y_{1,i}$ equals to one for an observation i that defaults, or zero otherwise, and n is the number of observations. In the RS algorithm, the smoothing terms $\mathbf{s}(\cdot)$ are modelled by Penalised B-splines (Eilers and Marx, 1996) because they enable smoothing parameter selection to be performed automatically by minimising the AIC $= -2L^p + 2N$, where N is the number of parameters in the model.

3.4.2 Marginal specification: balance model

As previously described, we separately model zero and non-zero balances and transform the latter one via the log transformation. The continuous non-zero log-transformed balance is denoted by Y_2, which takes a positive value. Gamma, Normal, Weibull, Inverse Gaussian, Log-normal, Logistic, and Gumbel distributions are considered as the candidates for the distribution of Y_2. The criteria utilised to identify the best model are the AIC and the BIC, Pearson correlation (discriminatory power), and Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) (predictive accuracy), each of which is evaluated on the validation set. The latter three measures are again computed as the mean across ten different subgroups of the validation set. The Logistic distribution shows the best performance and is hence selected (see Table 3.3). Figure 3.4 shows residual plots for all models. There is a noticeable deviation in the lower area, but overall, the plots reveal a decent model fit (except for the Gumbel distribution).

Therefore, we assume that Y_2 follows a Logistic distribution with CDF

$$F_{Y_2}(y_2|\mu,\sigma) = \frac{1}{1+e^{-\frac{y_2-\mu}{\sigma}}}, \qquad y_2 \in (-\infty,\infty), \tag{3.2}$$

Distribution	AIC	BIC	Correlation	RMSE	MAE
Gamma	72178	42947	0.9019	17443.24	7251.53
Normal	68825	69575	0.9020	17443.34	7265.05
Weibull	57170	57762	0.9006	17680.26	7595.57
Inverse Gaussian	74694	75457	0.9020	17430.91	7254.05
Log-Normal	74382	75155	0.9013	17484.37	7283.62
Logistic	47840	48550	0.9038	17131.43	7059.34
Gumbel	59233	59796	0.9015	17615.69	7609.78

TABLE 3.3: Performance measurements of the candidate marginal distributions for non-zero transformed balance Y_2, assessed on the validation dataset.

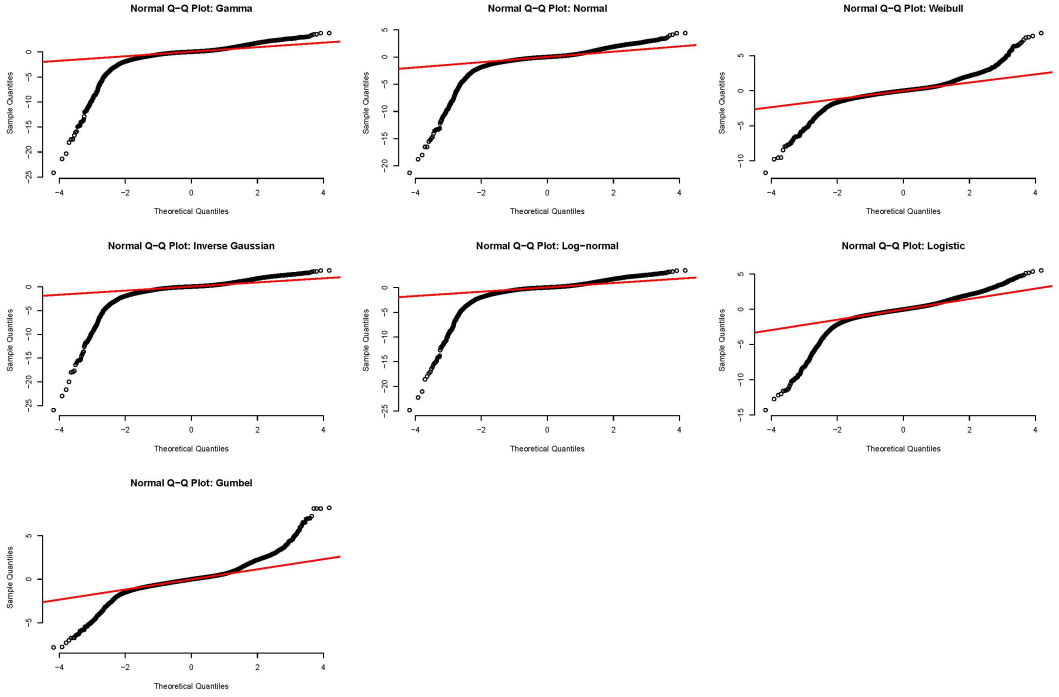


FIGURE 3.4: Residual plots for balance model.

and density

$$f_{Y_2}(y_2|\mu,\sigma) = \frac{1}{\sigma}\cdot\left[e^{-\frac{y_2-\mu}{\sigma}}\right]\cdot\left[1+e^{-\frac{y_2-\mu}{\sigma}}\right]^{-2}, \qquad y_2 \in (-\infty,\infty), \tag{3.3}$$

where $\mu \in (-\infty,\infty)$ and $\sigma > 0$ respectively denote the location and scale parameters of a Logistic distribution. The mean and variance are $E(Y_2) = \mu$ and $Var(Y_2) = \frac{\pi^2\sigma^2}{3}$, respectively. Since the Logistic distribution can take on non-positive values, we truncate the estimated mean to $\mu \in (0,\infty)$. The chosen model specifications for the μ and σ parameters are:

$$\begin{aligned}\mu =& \beta_0^\mu + \beta_1^\mu \text{age} + \beta_2^\mu \text{arr9} + s_3^\mu(\text{l}) + s_4^\mu(\text{b}) + s_5^\mu(\text{bsco}) + \\ & s_6^\mu(\text{paid.per9}) + s_7^\mu(\text{cu}) + s_8^\mu(\text{full.pay.per});\\ \log(\sigma) =& \beta_0^\sigma,\end{aligned} \tag{3.4}$$

where the logarithm link function for σ ensures its positivity. The additive smoothing functions are estimated by penalised B-splines with the likelihood function,

$$L = \prod_{i=1}^{n} f_{Y_{2,i}}(y_{2,i}|\mu_i, \sigma_i).$$

The fitting algorithm for the copula regression (shown in the next subsection) requires all model parameters to be computed simultaneously. Hence, introducing additional terms in the sigma equation (especially the non-parametric smoothing terms) would drastically increase the computational time and complicate the model. Also, since the main focus is on the mean level, we restrict σ to be related to the intercept term β_0^{σ} only. This makes the model less computationally expensive and simpler to implement in practice.

3.4.3 Copula specification

In this subsection, we specify how our proposed model is constructed from the two standalone marginal models under the CGAMLSS framework. The bivariate cumulative distribution function (CDF) of Y_1 and Y_2, $F_{Y_1,Y_2}(y_1, y_2)$, can be expressed as a combination of two marginal CDFs, F_{Y_1} and F_{Y_2}, with their dependence structure described by means of a copula (Sklar, 1959):

$$F_{Y_1,Y_2}(y_1, y_2) = C_\theta(F_{Y_1}(y_1), F_{Y_2}(y_2)), \tag{3.5}$$

where θ in the copula function C_θ is a (set of) dependence parameter(s) representing the interaction between the margins. Equation (3.5) illustrates how a copula enables a bivariate response vector to be flexibly defined by arbitrary marginals while allowing the dependence structure to be specified by the choice of a suitable copula. If the margins are both continuous, then the copula is unique. However, we study a mixed binary-continuous response, so the copula can no longer be uniquely determined. To overcome this challenge, we apply a latent variable representation for the binary regression model component. The binary variable Y_1 is assumed, without loss of generality, to be related to the (unobserved) continuous latent variable Y_1^* by defining $Y_1 = \mathbb{I}(Y_1^* > 0)$, where $\mathbb{I}(\cdot)$ is an indicator function. This can be expressed as $Y_1^* = \eta_\pi + U$, where η_π is the linear predictor of Y_1 specifying the success probability (see Equation (3.1)) and U is an error term with CDF, $F_U(u)$. Different distributions of the error term lead to different link functions in the binary regression. More specifically, Standard Logistic, Standard normal, and Gumbel distributions of U respectively lead to Logit, Probit, and Complementary log-log models of π. Earlier, the Logit specification was chosen for the standalone PD model, which corresponds to

$U \sim \text{Logistic}(0,1)$. Hence, $Y_1^* \sim \text{Logistic}(\eta_\pi, 1)$ and

$$F_{Y_1^*}(0) = \frac{1}{1+e^{\eta_\pi}}. \tag{3.6}$$

The CDFs of Y_1 and Y_1^* coincide at $y_1 = y_1^* = 0$ since
$P[Y_1 = 0] = P[Y_1 \leq 0] = F_{Y_1}(0) = F_{Y_1^*}(0) = P[Y_1^* \leq 0]$, and therefore, a mixed binary-continuous joint probability density function (PDF) can be written as follows:

$$\begin{aligned}
f_{Y_1,Y_2}(0,y_2) &= P[Y_1 = 0|Y_2 = y_2] \cdot f_{Y_2}(y_2) = P[Y_1^* \leq 0|Y_2 = y_2] \cdot f_{Y_2}(y_2) \\
&= \lim_{\epsilon \to 0} \frac{P[\{Y_1^* \leq 0\} \cap \{y_2 \leq Y_2 < y_2 + \epsilon\}]}{P[y_2 \leq Y_2 < y_2 + \epsilon]} \cdot f_{Y_2}(y_2) \\
&= \lim_{\epsilon \to 0} \frac{[F_{Y_1^*,Y_2}(0, y_2+\epsilon) - F_{Y_1^*,Y_2}(0,y_2)]/\epsilon}{[F_{Y_2}(y_2+\epsilon) - F_{Y_2}(y_2)]/\epsilon} \cdot f_{Y_2}(y_2) \\
&= \frac{\partial F_{Y_1^*,Y_2}(0,y_2)}{\partial y_2} \cdot \frac{1}{f_{Y_2}(y_2)} \cdot f_{Y_2}(y_2) \\
&= \frac{\partial C_\theta(u,v)}{\partial y_2}, \text{ where } \quad u = F_{Y_1^*}(0) \quad \text{and} \quad v = F_{Y_2}(y_2); \\
&= \frac{\partial C_\theta(u,v)}{\partial v} \cdot \frac{\partial v}{\partial y_2} = \frac{\partial C_\theta(u,v)}{\partial v} \cdot \frac{\partial F_{Y_2}(y_2)}{\partial y_2} = \frac{\partial C_\theta(u,v)}{\partial v} \cdot f_{Y_2}(y_2),
\end{aligned}$$

and

$$\begin{aligned}
f_{Y_1,Y_2}(1,y_2) &= f_{Y_2}(y_2) - f_{Y_1,Y_2}(0,y_2) \\
&= f_{Y_2}(y_2) \cdot \left(1 - \frac{\partial C_\theta(u,v)}{\partial v}\right), \text{ where } u = F_{Y_1^*}(0) \text{ and } v = F_{Y_2}(y_2).
\end{aligned}$$

In short, the joint PDF can be re-written as:

$$f_{Y_1,Y_2}(y_1,y_2) = [F_{1|2}(0|y_2)]^{1-y_1} \cdot [1 - F_{1|2}(0|y_2)]^{y_1} \cdot f_{Y_2}(y_2), \quad y_1 \in \{0,1\} \text{ and } y_2 > 0, \tag{3.7}$$

where $F_{1|2}(0|y_2) := P[Y_1 = 0|Y_2 = y_2] = \left(\frac{\partial C_\theta(F_{Y_1^*}(0), F_{Y_2}(y_2))}{\partial F_{Y_2}(y_2)}\right)$, $F_{Y_1^*}$ is the CDF of Y_1^* in Equation (3.6), F_{Y_2} is the CDF of Y_2 in Equation (3.2), and f_{Y_2} is the PDF of Y_2 in Equation (3.3). Equation (3.7) shows how the dependence, embedded in the copula function, is incorporated by a joint density function that is a simple product of a conditional probability of default given a level of balance and a balance density at that level.

Various forms of dependence between Y_1 and Y_2 can be selected through different copulas. We consider Gaussian, Frank, Farlie-Gumbel-Morgenstern (FGM), Clayton and Joe copulas, including their rotations, as potential candidates. Rotated copulas (Nelsen, 2006) can be constructed from the original copula. They allow for modelling any non-symmetric dependence structures that are not possible with the non-rotated versions, such as negative tail dependence. As shown in Table 3.4, the Frank, Joe, and

Copula Function	AIC	BIC	Average Kendall's Tau
Gaussian	107915	108979	0.2937
Frank	107881	108847	0.3540
FGM	108675	109608	0.1871
0°Clayton	108606	109590	0.2358
90°Clayton	665779	666468	-0.0016
180°Clayton	107423	108442	0.3555
0°Joe	107612	107669	0.3679
90°Joe	110124	110973	-0.00001
180°Joe	108617	109624	0.2368

TABLE 3.4: Performance measurements of the candidate copula functions, assessed on the validation dataset.

Copula Function	Brier score	AUROC	Correlation	RMSE	MAE	Q90
180°Clayton	0.1077	0.9219	0.9024	17325	7386	4970
0°Joe	0.1084	0.9215	0.9015	17413	7416	5000

TABLE 3.5: Predictive accuracy measurements of 180°Clayton and Joe copula functions, assessed on the validation dataset.

180°Clayton copulas are most supported by the AIC and BIC. The Frank copula implies a structure where dependence in the tail areas is weak, but it is strong in the middle of the marginal distributions. The Joe and rotated 180°Clayton, on the other hand, express strong right (upper) tail dependence but relatively weak dependence in the lower and middle area. As their properties are fundamentally different and with no clear winner among them thus far, we decide to build copula models using both the Frank and 180°Clayton copulas. The 180°Clayton was selected over Joe because it gives better predictive accuracy (see Table 3.5). The Frank copula is defined as

$$C_\theta^F(u,v) = -\frac{1}{\theta}\cdot\log\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right], \quad \theta\in(-\infty,\infty),$$

and the 180°Clayton copula is defined as

$$C_\theta^{C180}(u,v) = u+v-1+((1-u)^{-\theta}+(1-v)^{-\theta}-1)^{-1/\theta}, \quad \theta\in(0,\infty).$$

For both copulas, the higher the θ, the higher the dependence between the two margins.

The estimated conditional dependence between PD and balance given predictors' values, measured by Kendall's Tau, $\hat{\tau}$, can be expressed in terms of a copula as $\tau = 4\,E[C_\theta(U,V)]-1$ (Balakrishnan and Lai, 2009). For each copula, the average of this value (listed in the right-most column of Table 3.4) clearly demonstrates that PD and balance are positively correlated; its value under the copula rotations that imply negative dependence (i.e. 90°Clayton and 90°Joe) is nearly zero.

Together with Equation (3.1) and Equation (3.4), this gives the following copula model specification:

$$\begin{aligned} \text{logit}(\pi) &= \beta_0^\pi + \beta_1^\pi \text{age} + \cdots + s_6^\pi(\text{paid.per9}) + s_7^\pi(\text{cu}) + s_8^\pi(\text{full.pay.per}); \\ \mu &= \beta_0^\mu + \beta_1^\mu \text{age} + \cdots + s_6^\mu(\text{paid.per9}) + s_7^\mu(\text{cu}) + s_8^\mu(\text{full.pay.per}); \\ \log(\sigma) &= \beta_0^\sigma; \\ \text{G}(\theta) &= \beta_0^\theta + s_1^\theta(\text{l}) + s_2^\theta(\text{b}) + s_3^\theta(\text{bsco}) + s_4^\theta(\text{cu}), \end{aligned} \tag{3.8}$$

where $\text{G}(\theta) = \theta$ for the Frank copula and $\text{G}(\theta) = \log(\theta)$ for the 180°Clayton copula. Based on preliminary analyses, additive smoothing functions of a limited set of covariates are considered for the dependence equation, $\text{G}(\theta)$, to avoid overcomplicating the model. Using the derivation of Equation (3.7), the log-likelihood function is

$$l(\boldsymbol{\delta}) = \sum_{i=1}^{n} (1 - y_{1,i}) \log[F_{1|2}(0|y_{2,i})] + y_{1,i} \log[1 - F_{1|2}(0|y_{2,i})] + \log[f_2(y_{2,i})],$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}_\pi^T, \boldsymbol{\beta}_\mu^T, \boldsymbol{\beta}_\sigma^T, \boldsymbol{\beta}_\theta^T)^T$ is the vector of marginal and dependence parameters to be estimated from Equation (3.8). This likelihood extends the Heckman correction (Heckman, 1979) of sample selection bias to distributions other than Gaussian. Simultaneous parameter estimation, which accommodates the interplay of two responses, is accomplished by applying the trust region algorithm with integrated automatic multiple smoothing parameter selection (Marra and Radice, 2017a) and implemented in the R package GJRM (Marra and Radice, 2017b). This fitting algorithm has been proved to be fast and unbiased and can be done in a modular way, allowing any parametric marginal distributions or copula functions as long as their CDFs, PDFs and derivatives with respect to their parameters are acknowledged.

3.4.4 Probability of zero balance

We denote by ν the probability of account balance being zero. The modelling steps for ν are similar to those for PD, see the performance measures in Table 3.6 and Figure 3.5. The selected Logit model specification is:

$$\begin{aligned} \log\left(\frac{\nu}{1-\nu}\right) =& \beta_0^\nu + \beta_1^\nu \text{age} + \beta_2^\nu \text{arr9} + s_3^\nu(\text{l}) + s_4^\nu(\text{b}) + s_5^\nu(\text{bsco}) + \\ & s_6^\nu(\text{paid.per9}) + s_7^\nu(\text{cu}) + s_8^\nu(\text{full.pay.per}). \end{aligned}$$

The expectation of (zero and non-zero) balance unconditionally on default status, denoted as UB (short for unconditional balance), can thus be evaluated as:

$$E[\text{UB}] = (1 - \nu) \cdot E[Y_2]. \tag{3.9}$$

Link Function	AIC	BIC	Brier score	AUROC
Logit	12316	12715	0.0478	0.8445
Probit	12329	12729	0.0479	0.8445
Cloglog	12315	12716	0.0478	0.8444

TABLE 3.6: Performance measurements of the candidate marginal distributions for probability of zero balance, assessed on the validation dataset.

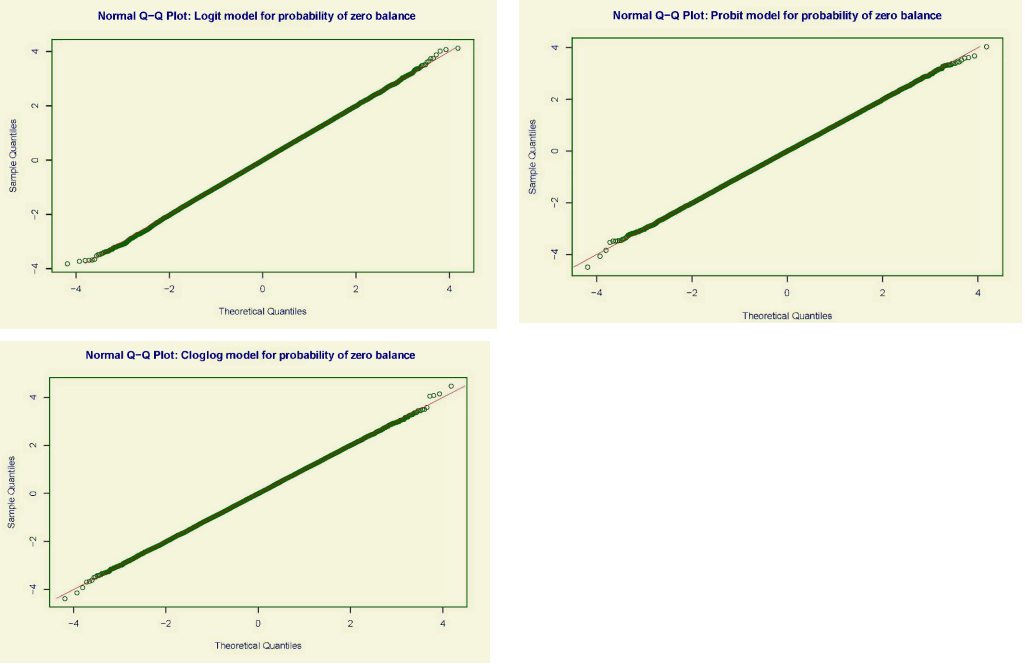


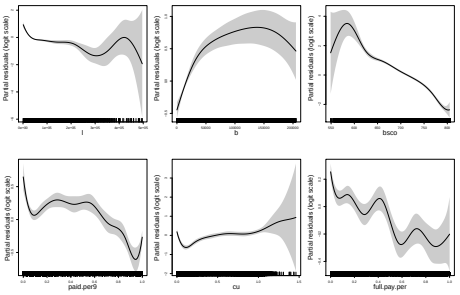
FIGURE 3.5: Residual plots for probability of zero balance model.

3.5 Analyses and results

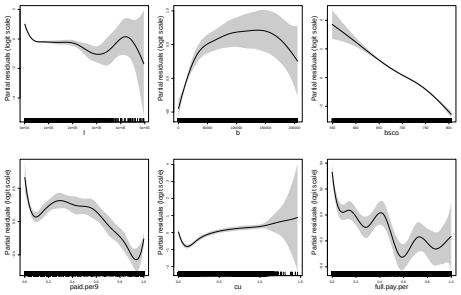
In this section, we will analyse and compare the results from the newly proposed CGAMLSS models against those from the standalone models. Significant relationships between the explanatory variables and the two responses, Y_1 and Y_2, will be inspected, as well as those for the dependence parameter(s) of the copulas. We will also report on the predictive ability of all models. In addition, the extent to which PD and balance affect each other will be examined by investigating the conditional distributions from the copula models. We conclude the section by analysing the impact of the dependence between PD and balance on expected loss.

3.5.1 Effects of covariates

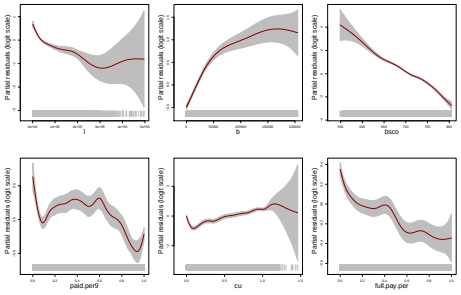
Figure 3.6 shows the effects of all model covariates on the probability of default. Results for the parametric terms are summarised in a table, whereas for the non-parametric smooth functions, partial residual plots depict how each explanatory



(a) Partial residual plots on logit scale for PD in the Frank copula model.



(b) Partial residual plots on logit scale for PD in the 180°Clayton copula model.



(c) Partial residual plots on logit scale for PD in the standalone PD model.

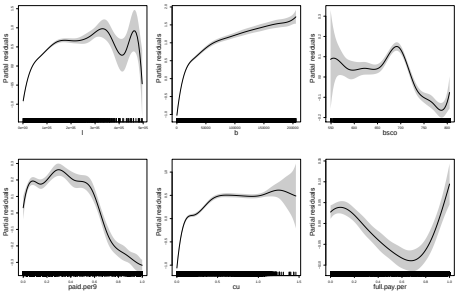
Model	Cop.Frank	Cop.C180	Ind.PD
(Intercept)	-0.3480***	-0.3801***	16.9809***
	(0.0310)	(0.0342)	(1.1847)
arr9 - Yes	0.1318***	0.2312***	0.2210***
	(0.0305)	(0.0344)	(0.0511)
age	-0.0011***	-0.0010***	-0.0011***
	(0.0002)	(0.0003)	(0.0003)

(d) Linear effects of explanatory variables on PD; standard errors are shown in parentheses and the significance level is identified at 10%(*), 5%(**) and 1%(***).

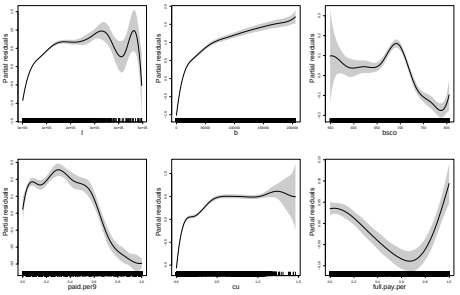
FIGURE 3.6: Effects of explanatory variables on PD for the standalone (Ind.PD) and the Frank (Cop.Frank) and 180°Clayton (Cop.C180) copula models.

variable influences the response assuming that the other covariates are fixed. The shaded areas in each plot indicate the precision of the estimates using 95% confidence intervals. Note that the waviness and widening confidence band near the lower or upper ends of the covariate range suggest some undersmoothing linked to there being fewer such observations. In line with expectations, credit card customers with a high rating and credit limit are at lower risk of default. So are customers with longer tenure, as well as those who previously paid back a higher proportion of their monthly balance or more often repaid the balance in full. In contrast, starting balance and credit utilisation are positively related to default risk; the more money drawn (either in absolute terms or as a percentage of the limit), the higher the risk of default. Also, having been in arrears recently tends to increase the likelihood of default.

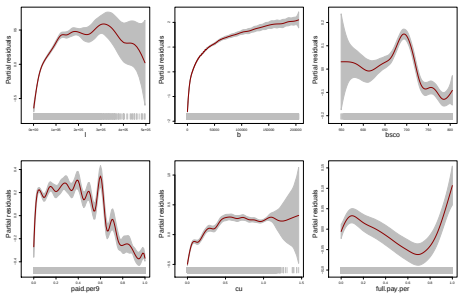
Secondly, Figure 3.7 displays the effects of covariates on the mean of (non-zero) balance. Overall, there is no significant difference in interpretation across the different models. Limit is positively associated with future balance. This is perfectly intuitive as customers with a higher limit are allowed to borrow more. Similarly, current levels of credit utilisation and balance are strongly linked to the subsequent balance, higher values implying larger future balance. Longer account tenure and prior arrears,



(a) Partial residual plots for the mean level of (non-zero) balance in the Frank copula model.



(b) Partial residual plots for the mean level of (non-zero) balance in the 180°Clayton copula model.



(c) Partial residual plots for the mean level of (non-zero) balance in the standalone balance model.

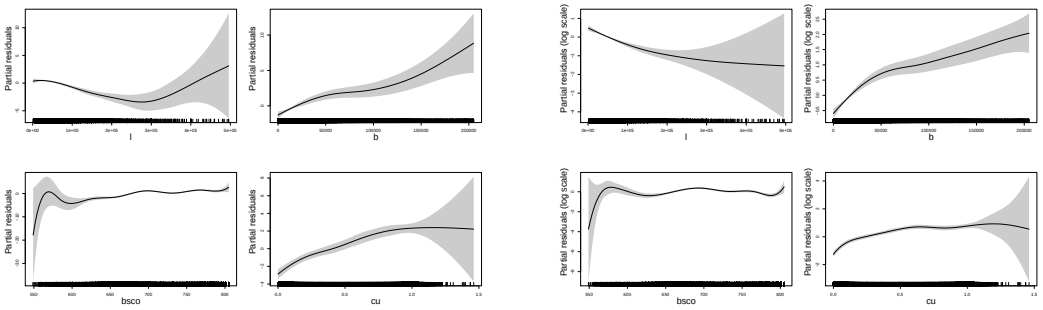
Model	Cop.Frank	Cop.C180	Ind.UB
(Intercept)	9.2253***	9.2112***	12.7000***
	(0.0073)	(0.0073)	(0.0298)
arr9 - Yes	-0.0783***	-0.0649***	-0.0833***
	(0.0093)	(0.0092)	(0.0085)
age	-0.0003***	-0.0003***	-0.0005***
	(7.17e-05)	(7.06e-05)	(7.05e-05)

(d) Linear effects of explanatory variables on the mean level of (non-zero) balance; standard errors are in parentheses and the level of significance is identified at 10%(*), 5%(**) and 1%(***).

FIGURE 3.7: Effects of explanatory variables on the mean level of (non-zero) balance for the standalone (Ind.UB) and the Frank (Cop.Frank) and 180°Clayton (Cop.C180) copula models.

however, tend to be associated with lower balance. Two novel insights are encountered. First, behavioural score and average paid percentage both have a concave effect on the response, indicating that it is more often that the borrowers of average credit quality or those who pay back between 10 to 60 percent of their monthly balance will borrow more. However, the convex effect plot for full payment percentage suggests that, holding other factors constant, it is those that either rarely or most of the time pay back the owed money in full, that tend to have a higher future balance. These non-monotonic effects demonstrate the potential benefits of introducing non-linear effects into the models.

Next, we consider the association between the two responses. The average estimated conditional Kendall's tau, $\hat{\tau}$, for the Frank (0.37 with 95% CI (0.33,0.42)) and 180°Clayton (0.38 with 95% CI (0.34,0.41)) copula models is positive, as expected. However, the copula models allow us to investigate whether this positive dependence is modified by the covariates. Figure 3.8 displays the effects of covariates on the dependence parameter, θ. Interestingly, the dependence between default risk and future balance is stronger for borrowers with a higher current balance or utilisation,



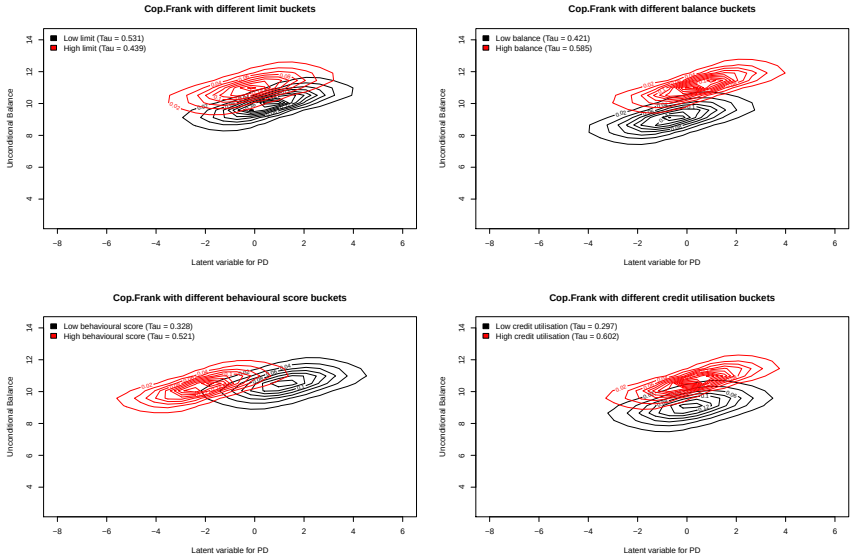
(a) Partial residual plots for dependence parameter in the Frank copula model.

(b) Partial residual plots on log scale for dependence parameter in the 180°Clayton copula model.

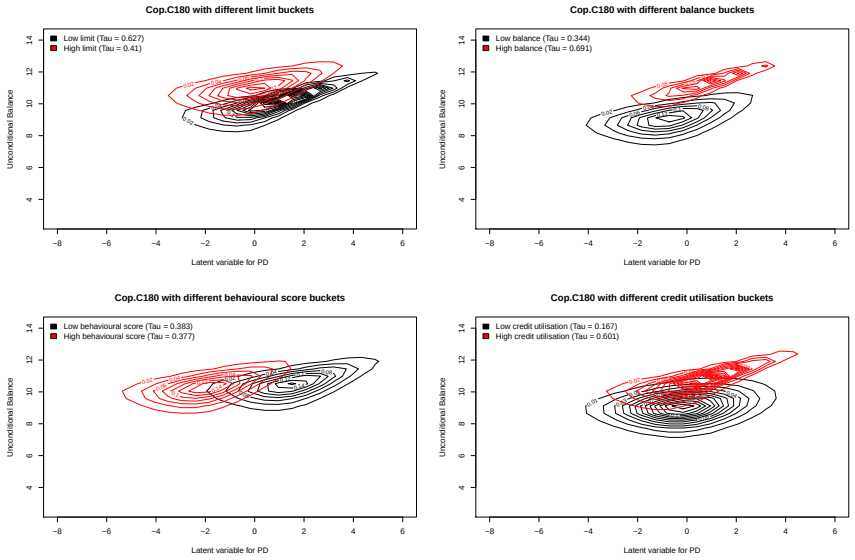
FIGURE 3.8: Effects of explanatory variables on dependence parameter for the Frank and 180°Clayton copula models.

and further varies with the credit limit, for the most part decreasing. These insights may prove useful in practice as banks could expect higher levels of borrowing from accounts with those characteristics in the event they default; this in turn suggests more capital is needed to cope with the stronger adverse dependency in this account segment. Another benefit of the copula approach over building standalone models lies in the additional finding that the dependence of PD and balance is not always appropriately modelled by a symmetrically linear structure. Our analysis shows that they correlate more heavily in the middle area (under Frank) or the upper tail (under 180°Clayton).

Lastly, to visually gauge the impact of a given input variable on the joint PDF of latent variable Y_1^* and balance, contour plots provide a helpful tool. Figure 3.9 displays a selection of contour plots for four different covariates: limit, current balance, behavioural score and utilisation. They help us to inspect the effect of each on three parameters of interest, namely π, μ and Kendall's τ, by categorising the covariate into two groups (low vs. high value) whilst keeping the others fixed at their mean (for continuous variables) or mode (discrete variables). For instance, under the Frank copula, a higher current balance (see the upper right plot in Figure 3.9a) leads to a higher PD (the cloud shifts to the right), a higher future balance (upward shift), and a higher association (higher Kendall's Tau). Hence, we can identify which covariates most influence each respective parameter. Under both copula models, credit rating is thus found to be the variable with the largest impact on PD, future balance appears most strongly influenced by current balance, while it is credit utilisation that has the largest effect on the conditional dependence between PD and balance.



(a) Frank Copula Model.



(b) 180°Clayton Copula Model.

FIGURE 3.9: Contour plots for the joint PDF of the latent variable Y_1^* and (non-zero) balance for different levels of limit (upper left), current balance (upper right), behavioural score (lower left), and credit utilisation (lower right), where the other predictors are fixed at their mean or mode levels. Estimated conditional dependencies, $\hat{\tau}$, are listed in the upper left corner of each plot.

3.5.2 Predictive performance

The hold-out performance measurements (averaged over ten different subgroups of the test dataset) for the point estimates from all PD and balance models are shown in Table 3.7.

The low Brier scores and high AUROC values in the left panel indicate good performance for all PD models. The PD estimate from the standalone model appears to perform better than that of the copula models but the winning margin is very small.

Model	Brier score	AUROC
Cop.Frank	0.1086	0.9226
Cop.C180	0.1069	0.9229
Ind.PD	0.1047	0.9244

(A) PD models.

Model	Correlation	RMSE	MAE	Q90
Cop.Frank	0.9159	16139.89	7054.01	4668.36
Cop.C180	0.9160	16158.19	7025.90	4709.75
Ind.UB	0.9157	16171.23	7062.61	4690.56

(B) Balance models.

TABLE 3.7: Performance measurements assessed on the test set.

Similarly, the discriminatory power and predictive accuracy differences for the standalone and copula models for balance are negligible (see the right panel). The same goes for the 0.9 quantile loss (Q90), which penalises underestimation more heavily and, hence, is a good measure for assessing the conservativeness of a risk estimate. Note that the α quantile loss function is defined as $\sum_{i;y_i<\hat{y}_i}(\alpha-1)\cdot(y_i-\hat{y}_i)+\sum_{i;y_i\geq\hat{y}_i}\alpha\cdot(y_i-\hat{y}_i)$, where y_i and $\hat{y}_i$ are true and predicted balance values, respectively.

These results imply that the three approaches are equally competitive in terms of the account-level point estimates that they produce for future balance; however, from a risk perspective, the goodness-of-fit of the conditional distributions produced by each approach may be of greater interest, as well as the copula models' ability to model dependence between default risk and balance.

3.5.3 Conditional probability, density, and expectation

In the standalone models, the distributions of PD and balance depend on a set of (partially shared) explanatory variables but, after accounting for those, they are assumed independent from each other. In the copula models, however, the distributions of the two responses are also conditionally dependent on one another. In other words, the value for PD directly depends on the value for balance and vice versa. To better understand the dependence captured by the copula models, we will consider the resulting conditional default risk in each quantile of balance and, vice versa, the distribution of expected balance for defaults versus non-defaults. Note that, as the subsequent analysis thus assumes knowledge of the other response outcome, it is not meant for assessing any prediction applications but to better understand the explanatory power of the models. Following Equation (3.7), we calculate the conditional PD for a given (non-zero) balance amount as:

$$P[Y_1=1|Y_2=y_2]=\frac{f_{Y_1,Y_2}(1,y_2)}{f_{Y_2}(y_2)}=1-F_{1|2}(0|y_2).$$

We also obtain the conditional density and conditional expectation of (non-zero) balance given default status, as follows:

$$f_{Y_2}(y_2|Y_1 = y_1) = \frac{[F_{1|2}(0|y_2)]^{1-y_1} \cdot [1 - F_{1|2}(0|y_2)]^{y_1} \cdot f_{Y_2}(y_2)}{P[Y_1 = y_1]};$$
$$E[Y_2|Y_1 = y_1] = \int_0^\infty x \cdot f_{Y_2}(x|Y_1 = y_1)\, dx.$$

Hence, the conditional expectation of balance is
$E[\text{UB}|Y_1 = y_1] = (1 - \nu) \cdot E[Y_2|Y_1 = y_1]$, where ν is the probability of balance being zero.

Firstly, for every defaulted and non-defaulted account observation in the test set, we consider: the actual value of (future) balance, UB; $E[\text{UB}|Y_1]$, i.e. the model estimate conditional on the observed default outcome, according to each copula model; and the (unconditional) $E[\text{UB}]$ from the standalone model. Secondly, the observations in the test set are sorted according to UB and split into quantiles. In each quantile interval, $(q_i, q_{i+1}), i = 1, \ldots, 999$, we then calculate: the (actual) empirical proportion of defaults; the mean conditional probability of default estimated by each copula model for the interval's midpoint, $P[Y_1 = 1|Y_2 = (q_i + q_{i+1})/2]$; and the mean estimated (unconditional) $P[Y_1 = 1]$ from the standalone model. Figure 3.10 plots these (estimated or observed) default rates (y-axis) against each balance quantile (x-axis).

In Figure 3.10, the empirical default rate curve (Actual) shows the (non-linear) relationship that exists between (future) balance and default risk. Despite being fitted without taking this dependence into account explicitly, the standalone model for PD (Ind.PD) is already capable of capturing a fair proportion of the co-movement between PD and balance. This is due to covariates being included in both marginal models that simultaneously influence PD and balance. For instance, in section 3.5.1, higher levels of current balance and credit utilisation were found to increase both the risk of default and the future balance (see Figure 3.6 and Figure 3.7). However, the copula models have the added ability to capture any remaining stochastic dependence that cannot be explained by those observable shared covariates. This explains why, with the added knowledge of balance, the conditional PD curves for the two copula models (Cop.Frank and Cop.C180) move further towards the actual default rates.

Secondly, to better understand how well the models capture the difference in balance between defaults and non-defaults, Figure 3.11 and Figure 3.12 show the mean and density plot, respectively, of actual and the estimated expected value of (conditional) balance. Note that, since balance cannot take negative values, we fit the probability density function by zero-truncated kernel density estimation with a Gaussian kernel and weight $w(x) = \frac{1}{1-\Phi_{x,h}(0)}$, where h is a bandwidth and Φ is the cumulative distribution function of a Gaussian distribution with mean x and standard deviation h. The objective is to truncate the density on the negative side at zero and up-weight

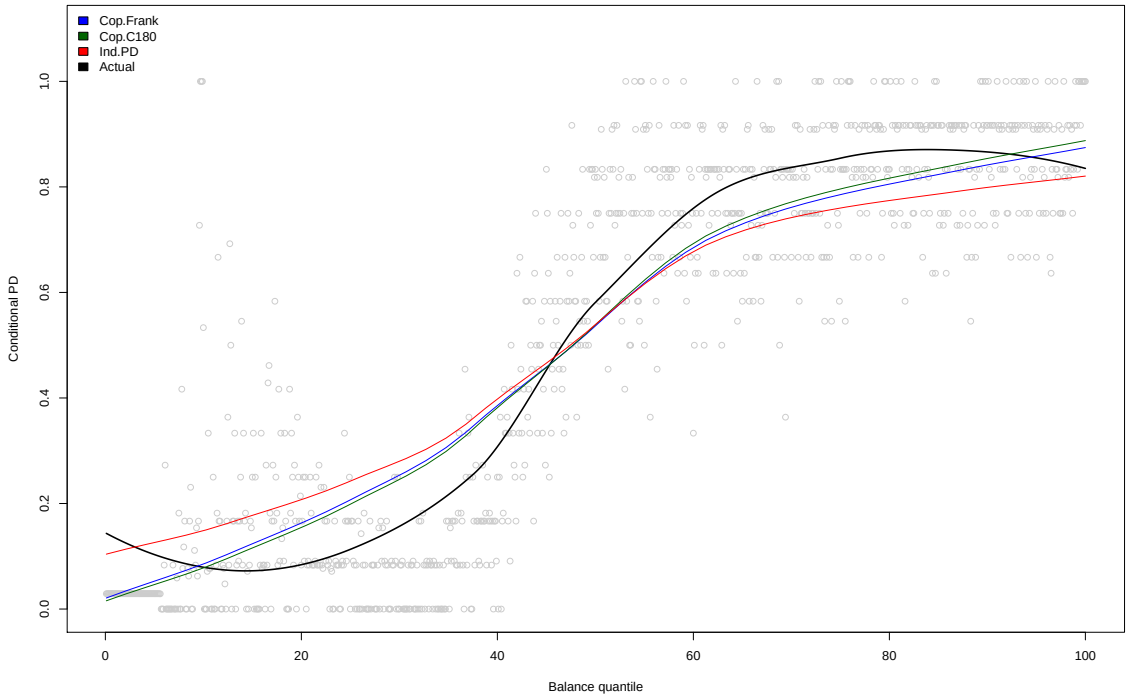
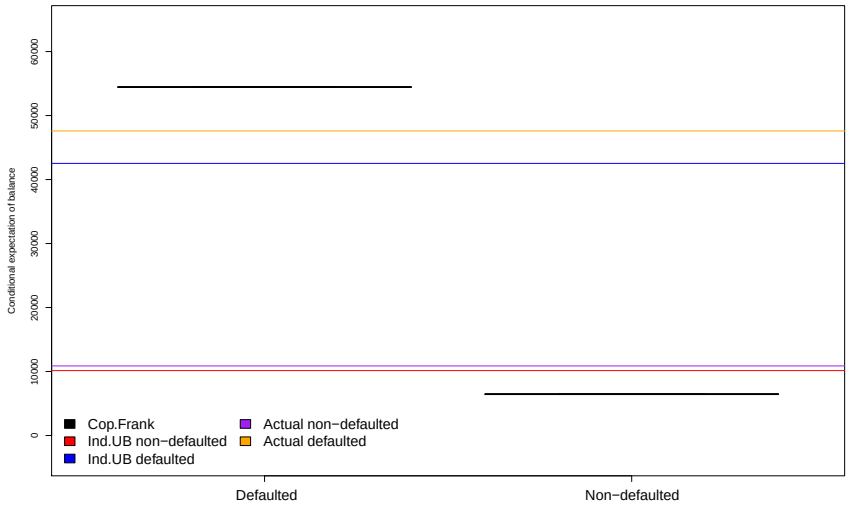
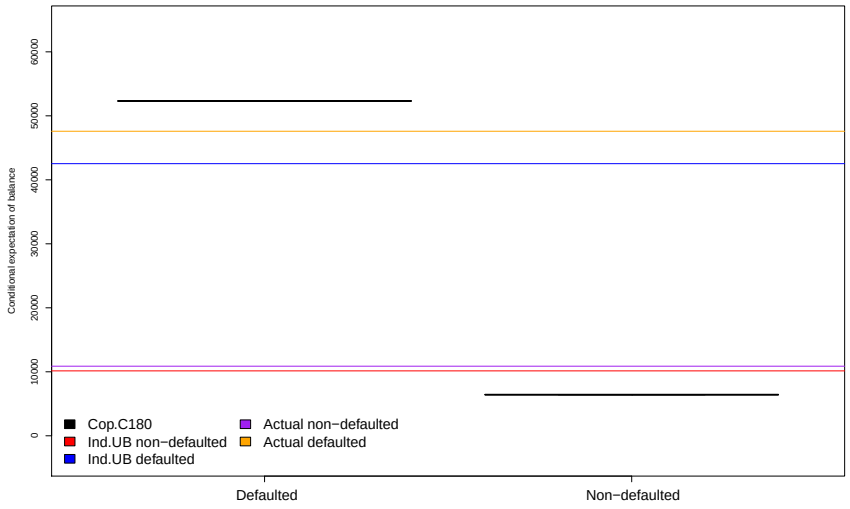


FIGURE 3.10: Average (conditional) PD given the level of balance quantile, using the locally weighted smoothing line technique and assessed on separated test set ranging from low to high quantiles of future balance. Small dots in the background represent the (actual) empirical proportion of defaults for each quantile interval.

the data that are close to zero. From these figures, it is evident that the actual balance of defaulted accounts tends to be higher (on average) than that of the non-defaults and exhibits a heavy positive tail. Taking a future estimate of balance based on all accounts and ignoring this dependence would thus lead to underestimating the balance at default (or EAD) and, hence, underestimating the expected loss. Due to its covariates, the standalone model for balance does, however, partially capture such differences in balance, similarly to what was observed in the preceding analysis. Importantly, we again see the copula models being able to capture residual dependencies by the use of a copula function, shifting the mean model estimate for balance further up (down) for the group of defaults (non-defaults), respectively (see Figure 3.11). Also, the estimated balance for defaulted accounts shows a heavier tail with the copula models than with the standalone model (see Figure 3.12); if anything, they appear to now somewhat overestimate (rather than underestimate) the dependence between the two responses (hence, producing a more conservative EAD estimate). Among the two copula models, the Frank copula model gives the most conservative estimates in terms of mean expected conditional balance given default (see Figure 3.11).



(a) Frank Copula Model.

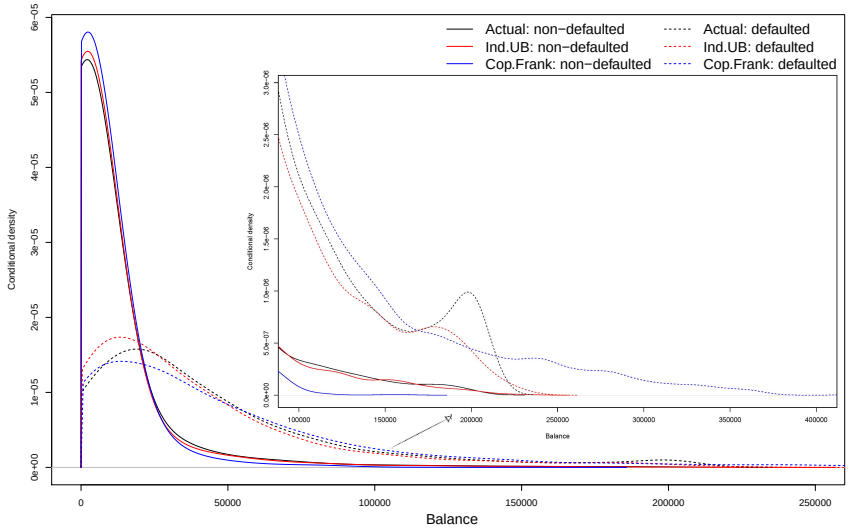


(b) 180°Clayton Copula Model.

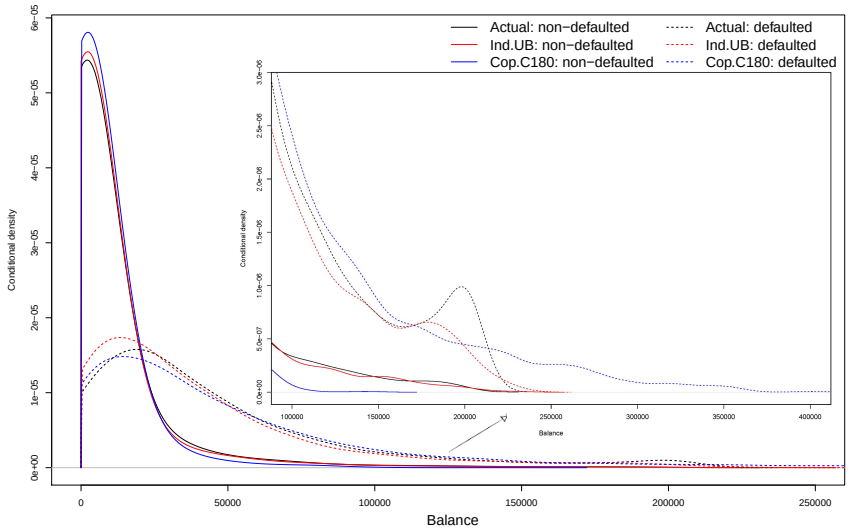
FIGURE 3.11: The mean of the expected value of (conditional) balance given default status, assessed on separated defaulted and non-defaulted accounts in the test set.

3.5.4 Expected loss estimation

One of the hypothesised advantages of the proposed copula models is that they may help avoid potential sample selection problems (associated with the standard Basel practice of building an EAD model on just the subsample of defaulted accounts and then applying it to estimate EAD for all accounts), whilst keeping the ability of producing sufficiently conservative estimates. To test this, we will compare the expected loss amounts produced by our copula models against those produced by the aforementioned approach and the standalone models.



(a) Frank Copula Model.



(b) 180°Clayton Copula Model.

FIGURE 3.12: Density plots for the expected value of (conditional) balance given default status, assessed on separated defaulted and non-defaulted accounts in the test set; the right tail area is magnified.

Under the Basel framework, the credit loss associated with an account is seen as the product of three risk parameters,

$$\text{Loss}_B = Y_1 \cdot \text{LGD} \cdot \text{EAD}, \tag{3.10}$$

where Y_1 is the default status (i.e. 0 or 1), LGD is the Loss Given Default and EAD is the Exposure At Default. For the sake of simplicity, and as the paper has not included models for LGD, we shall from here on assume LGD to be fixed at one; i.e., we do not consider any recovery or collection process after default. Note that the analysis could

easily be extended if such data were available.

Firstly, as is common practice in Basel models, we build another model for EAD using only the defaulted accounts from the training set, thereby following a similar approach to that used for fitting the standalone balance model earlier (see subsection 3.4.2). Since the risk weight functions in the Accords do not directly consider adverse dependencies between default and exposure (other than through the use of downturn estimates as inputs), the Basel regulatory expected loss is

$$E[\text{Loss}_\text{B}] = E[Y_1] \cdot 1 \cdot E[\text{EAD}] = \text{PD} \cdot E[\text{EAD}],$$

where PD and EAD are modelled independently.

Secondly, for the purpose of comparison, we consider using all available data (defaulted and non-defaulted), first assuming that Y_1 and UB (balance) are independent. Thus, under the independent standalone framework, the expected loss would be

$$E[\text{Loss}_\text{I}] = E[Y_1] \cdot 1 \cdot E[\text{UB}] = \text{PD} \cdot E[\text{UB}],$$

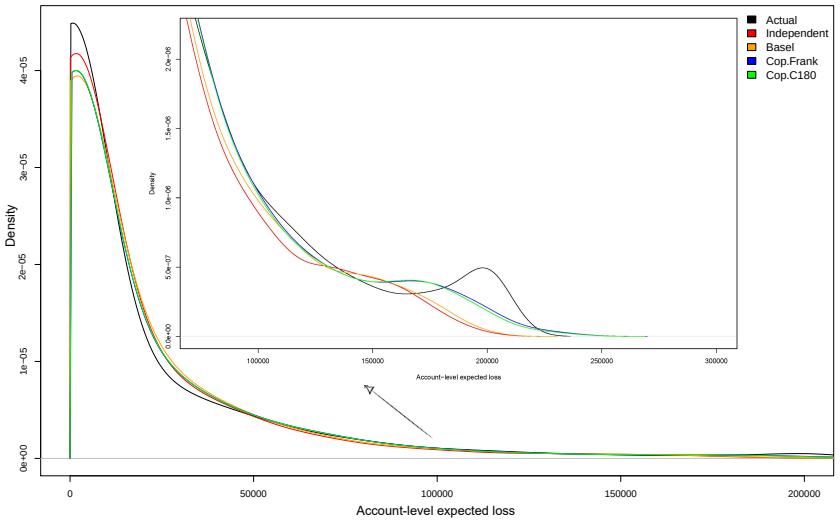
where PD and UB are modelled independently using the standalone models.

Thirdly, if the dependence between the two responses is instead taken into account, the expected loss, now under the copula framework, would be

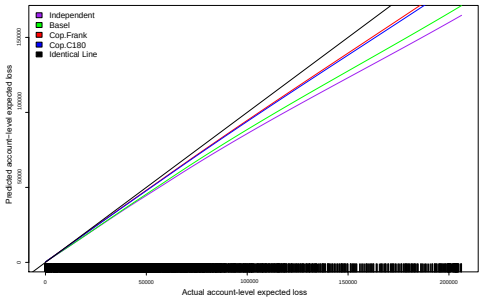
$$\begin{aligned} E[\text{Loss}_\text{C}] = E[Y_1 \cdot 1 \cdot \text{UB}] &= \int_0^\infty f_{Y_1,\text{UB}}(1,x) \cdot 1 \cdot x \, dx \\ &= \int_0^\infty f_{Y_1}(1) \cdot f_{\text{UB}}(x|Y_1 = 1) \cdot x \, dx = \text{PD} \cdot E[\text{UB}|Y_1 = 1], \end{aligned} \tag{3.11}$$

where PD and UB are modelled simultaneously using the copula models.

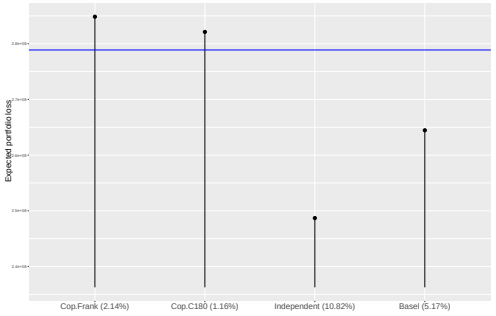
Figure 3.13 compares the estimated expected loss under these three modelling assumptions (i.e. for the Basel, independent and copula frameworks) against the actual values. As illustrated in Figure 3.13a, the estimation under the standalone scheme performs better in the lower loss space whereas the predictions under the copula framework are more accurate in the right tail area. As the right-tail area has greater implications for loss calculations, this suggests that the copula approach may be preferable. The same is suggested by the calibration plot in Figure 3.13b, which shows that, taking into account the dependence between PD and balance using the copula approaches leads to better calibrated account loss estimates where the actual loss is higher. As Table 3.8 demonstrates, this leads to better predictive accuracy at the account level (lower MAE) for Cop.Frank and Cop.C180, compared to the other two methods. Furthermore, the copula models do so by providing more conservative predictions (cf. their Q90, which is considerably lower). Note that MAE is preferable to RMSE as a monetary loss measure since the latter returns a square unit which is more difficult to interpret. It is also expected to see higher values of RMSE for the



(a) Density plots of account-level expected loss.



(b) Calibration plots comparing actual and predicted account-level expected loss. The locally weighted smoothing line technique is used and the identical line is represented in black.



(c) Comparison of expected portfolio loss; the actual value is shown at the horizontal line.

FIGURE 3.13: Loss analyses, assessed on all accounts in the test dataset.

copula models because they: (1) emphasise the dependency between PD and balance resulting in a higher prediction for extreme losses; and thus (2) are penalised by RMSE much more heavily than MAE due to those extreme errors. Although they are close, out of the two copula methods, it is the Frank copula that captures stronger dependence and hence produces the most conservative loss estimates.

Model	Correlation	MAE	RMSE	Q90
Cop.Frank	0.8475	8075.70	21027.24	3833.61
Cop.C180	0.8475	8083.61	20956.27	3930.83
Independent	0.8476	8637.42	20903.32	5348.73
Basel	0.8501	8644.12	20603.64	4814.21

TABLE 3.8: Performance measurements (averaged over ten different subgroups of the hold-out test set) for account-level expected loss.

In Figure 3.13c, the loss analysis is extended to the portfolio level. When adding up the account-level estimates to produce an estimate of total portfolio loss (for the test set), the 180°Clayton copula model produces the best performance with a small absolute percentage error of 1.16%, followed by the Frank copula model (which overestimates portfolio losses by 2.14%). This is consistent with the previous results showing a stronger dependence effect under the Frank copula. Alternatively, calculating portfolio loss without considering the dependence (cf. Independent) or sample selection bias (cf. Basel) prompts a substantially worse underestimation (by 10.82% and 5.17%, respectively), suggesting substantial capital shortfalls.

We conclude that the CGAMLSS approach produces more accurate and sufficiently conservative expected loss estimates, at both the individual account level as well as the portfolio level. This suggests that the correlation between the two standalone model responses induced by their shared covariates alone is not sufficient to capture the full extent of their dependence; there exist remaining stochastic dependencies from non-recorded or unobservable factors that could only be captured by the copula approach.

3.6 Conclusions and future research

In this paper, we have proposed a novel approach for modelling PD, balance and their dependence structure simultaneously, by applying the CGAMLSS framework. Using this framework, a bivariate distribution could be flexibly constructed from two marginal GAMLSS responses and a suitable copula. This approach addresses the potential problem of sample selection bias identified in the EAD literature, by including both defaulted and non-defaulted account information in the modelling. For a large dataset of credit card accounts, our analysis shows that the proposed copula models provide more accurate and conservative expected loss estimates, exhibiting a heavy tail that is the result of the correlation between PD and credit card balance. In addition, we have demonstrated that by, instead, ignoring such dependence or by allowing sample selection bias, loss could be severely underestimated, potentially leading to capital shortfalls.

The results reveal strong and positive dependence between PD and balance, even after accounting for observable covariates, either in the middle (under Frank) or upper tail (under 180°Clayton) area of the marginal distributions. Accounts with higher default likelihood tend to end up having a higher card balance; also, the distribution of future balance shows a heavier tail for accounts that are more likely to default. In addition, we identify a series of significant risk factors: credit rating has the largest impact on PD, future balance is most strongly influenced by current balance, and their co-dependence is most affected by credit utilisation.

A future avenue of research is to extend the proposed model so that PD-LGD and LGD-EAD dependencies are considered as well. This would provide further insights on how the interaction between all three Basel IRB parameters could affect the expected portfolio loss and, hence, the capital requirement. We conjecture that by modelling PD-LGD-EAD together, with their interrelationship, a more conservative portfolio loss estimate, with a heavier tail, could be expected.

Chapter 4

Modelling Credit Card Exposure At Default Using Vine Copula Quantile Regression

Suttisak Wattanawongwan [a,1], Christophe Mues [b], Ramin Okhrati [c], Taufiq Choudhry [b], Mee Chi So [b]

[a] School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[b] Southampton Business School, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

[c] Institute of Finance and Technology, University College London, London, WC1E 6BT, UK

[1] Corresponding author
Email address : S.Wattanawongwan@soton.ac.uk

Abstract

To model the Exposure At Default (EAD) of revolving credit facilities, such as credit cards, most of the research thus far has employed point estimation approaches, focusing on the central tendency of the outcomes. However, such approaches may have difficulties coping with the high variance of EAD datasets and their non-normal empirical distributions, whilst information on extreme quantiles, rather than the mean, can have greater implications in practice. Also, many of the input variables used in EAD models are strongly correlated, which further complicates model

building. This paper, therefore, proposes the vine copula-based quantile regression model, an interval estimation approach, to model the entire distribution of EAD and predict its conditional mean and quantiles. This novel methodology addresses several drawbacks of classical quantile regression including quantile crossing and multicollinearity and allows the multi-dimensional dependencies between all variables in any EAD dataset to be modelled by a suitable series of (either parametric or non-parametric) pair-copulas. Using a large dataset of credit card accounts, our empirical analysis shows that the proposed non-parametric model provides better point and interval EAD estimates and more accurately reflects its actual distribution compared to other models.

4.1 Introduction

Under the Advanced Internal Ratings-Based (A-IRB) approach, the Basel II and III Accords allow authorised banks to use their own methods to establish risk-sensitive capital requirements as a function of different credit risk parameters. The three key parameters are: Probability of Default (PD), i.e. the likelihood that a borrower will default or be unable to fulfil their repayment obligations; Exposure At Default (EAD), i.e. the expected gross exposure of the borrower at the time of default; and Loss Given Default (LGD), i.e. the percentage of this amount that the lender would not be able to recover. In credit risk, PD and LGD have thus far been the main centre of attention, whereas EAD has been studied far less. This paper focuses on the latter.

In the literature, the proposed statistical models for EAD tend to focus on producing accurate point estimates for the central tendency of the outcomes, i.e. the conditional mean. Unlike interval estimates, point estimates may, however, prove less useful given the non-normality and high variance encountered in EAD data (see e.g. Thackham and Ma (2018) and Leow and Crook (2016)). Furthermore, when estimating potential monetary losses in risk management or the capital required to absorb them, the most useful information lies in extreme risks in the right tail area, i.e. higher quantiles. Therefore, to better understand the EAD distribution, it is important to consider the estimation of EAD at different quantiles (e.g. 99% value-at-risk), rather than solely at the mean level. In this paper, we apply two interval estimation models to EAD modelling: linear quantile regression (Koenker and Bassett, 1978) and D-vine copula-based quantile regression (Kraus and Czado, 2017; Schallhorn et al., 2017).

The first of these two approaches is well known and frequently used in predicting conditional quantiles of a response variable given the values of covariates. It is robust to outliers and heteroscedasticity and makes no assumptions about the response distribution. However, two common pitfalls of using the method are the problem of quantile crossing (i.e. the crossing of regression lines of different quantile levels,

causing interpretation difficulties) and its ability to cope with correlations between the covariates. The latter is of particular interest because many of the input variables commonly used in EAD models are strongly correlated with each other. For instance, Tong et al. (2016) and Leow and Crook (2016) incorporated both current credit limit and card balance in the models, which can lead to multicollinearity problems and interpretation issues with the estimated coefficients. In contrast, the D-vine copula-based quantile regression approach will allow us to tackle those issues, by modelling such dependencies between the explanatory variables through a series of pair-copulas.

Whereas much of the credit risk literature on EAD modelling has analysed corporate credit (Gürtler et al., 2018), our models are fitted to a large dataset of credit card defaults, provided by a Hong Kong retail lender. For most A-IRB banks, credit cards account for the largest number of defaults, which are often scarce in practice among revolving line products (Qi, 2009). This enables building more advanced statistical models based on the available default data.

In the analysis, we will identify to what extent the magnitude of predictor effects varies for different sections of the EAD distribution, i.e. at the mean and different quantile levels. This is useful to assess risk drivers of the tail risk of EAD. In addition to examining the relationships between EAD and the covariates, we will also explicitly consider correlations between the covariates themselves, by utilising vine copulas. We will implement the proposed model using the R package vinereg (Nagler and Kraus, 2019), which provides various options of copula families including parametric and non-parametric ones. To empirically test the effectiveness of the two proposed quantile models in the context of EAD modelling, we benchmark them against an OLS model. In so doing, we will show how the proposed approaches lead to better point and interval estimates.

The rest of the paper is presented as follows. The relevant literature is reviewed under Section 4.2, from which the main contributions of the paper are then identified. Section 4.3 explains the data and variables used, and Section 4.4 provides a brief description of vine copulas. Section 4.5 illustrates how the statistical models are constructed. The results are analysed in Section 4.6. Section 4.7 concludes.

4.2 Literature review

Our review of the literature will begin by reviewing some of the existing work on EAD modelling and then turn its attention to the methods proposed in the paper. At the end, we will list the main contributions of our work.

4.2.1 EAD modelling

For revolving line products including credit cards, the Basel Accords have suggested an indirect way of calculating EAD by evaluating the Credit Conversion Factor (CCF), i.e. the proportion of the undrawn amount that will be drawn by the time of default (Valvonis, 2008). Despite its popularity, such approach has several drawbacks. First, the empirical CCF distribution does not conform to several statistical distributions and is highly bimodal. Second, its estimates must be restricted to the [0,1] range. Third, the modelling may struggle to cope with the contracting denominator when the current drawn amount is already close to the limit. For those reasons, alternative methods have been put forward, which include modelling EAD directly, as a monetary amount (as opposed to a ratio).

For example, Thackham and Ma (2018) modelled EAD directly (albeit for corporate revolving facilities) and captured its relationship with the credit limit by considering a three-component model, conditioning the EAD target variable on whether the limit was lowered or not. They used Ordinary Least Squares (OLS) regression to predict the mean level of EAD. Tong et al. (2016) applied a zero-adjusted gamma distribution under the Generalised Additive Models for Location, Scale and Shape (GAMLSS) framework (Stasinopoulos et al., 2017), to capture the EAD distribution observed in a dataset of UK credit card defaults. The proposed model was shown to outperform several benchmark models (including CCF ones) in terms of the mean level of EAD. Hon and Bellotti (2016) forecast drawn credit card balances not only at default time but at every time step, unconditional on a default event occurring. Different methods were compared, including OLS, two-stage regression, and random effects panel models. Similarly, Leow and Crook (2016) constructed a mixture model that considers the entire time period up to default. Rather than the balance, they proposed modelling the limit under the scenario that an account's borrowing hits the credit limit at least once in the race to default. None of these methods explicitly studied interval estimates, however, although Tong et al. (2016) did model a dispersion parameter.

4.2.2 Quantile regression

The prediction of conditional quantiles of the response variable given the values of covariates has found a variety of applications in many domains, including finance, where it became a fundamental instrument for risk management (Kraus and Czado, 2017; Adrian and Brunnermeier, 2016; Bouyé and Salmon, 2009). Linear quantile regression, established by Koenker and Bassett (1978), is a well-known method for estimating the conditional quantiles. For example, in the consumer credit risk setting, Somers and Whittaker (2007) previously used quantile regression to model the value

distribution of repossessed properties, which was then used to produce loss given default estimates for mortgage loans.

Modelling EAD with the use of quantile regression would be beneficial in several respects. Firstly, it considers the entire conditional distribution of EAD, which enables the estimation of conditional quantiles and confidence intervals, reveals any potential heavy tails and skewness, and allows for the shape of the distribution to depend on the covariate values. Secondly, it provides a comprehensive picture of the predictor effects on different quantiles of the EAD distribution, not only on the mean level. Thirdly, quantile regression is robust to outliers, which are often encountered in EAD data. Lastly, unlike least squares regression, it does not require the assumptions of a specific parametric distribution or constant variance for the response, making it an attractive alternative to account for heteroscedasticity (Niemierko et al., 2019).

However, classical (linear or non-linear) quantile regression has been criticised for several pitfalls. Kraus and Czado (2017) and Bernard and Czado (2015) highlighted the problem of quantile crossing; this is where the regression lines of different quantile levels (with distinctive slopes) cross each other, thus causing interpretation problems. The method also suffers from multicollinearity, i.e. strong correlation between the explanatory variables, making the estimated regression coefficients harder to interpret and unstable with large variances (Bager, 2018). This issue is highly relevant to EAD and other consumer credit data, since the variables in these settings are often associated with each other, either directly or indirectly; for instance, banks often actively manage the borrower's limit amount according to their balance expenditure. In addition, quantile regression does not acknowledge multivariate dependencies between the variables of interest, which are needed for credit portfolio risk modelling (Geidosch and Fischer, 2016). Conventional correlation analysis, assuming the popular, yet restrictive, multivariate Gaussian distribution, is not appropriate to investigate such underlying dependencies, because it cannot accommodate a non-linear and asymmetric structure, which has proven important in financial applications (see, e.g., Aas et al. (2009), Moreira (2010) and Geidosch and Fischer (2016)).

4.2.3 Copulas

Copulas are a more appropriate method to model complex dependence patterns. They allow a multivariate distribution to be jointly constructed from arbitrary univariate distributions, using an appropriate copula function. An attractive feature of copulas is that the functional forms of a copula and its components (marginal CDFs) can be selected independently. This gives them a key advantage over a conventional parametric specification (e.g. multivariate Gaussian) where the joint and marginal distributions must be known a priori. Moreover, various dependence structures

between individual variables can be captured by different copula specifications. For instance, the Clayton copula reflects lower tail dependence, whilst the Gumbel copula allows for stronger dependence in the upper tail area. The Student-t copula is both lower- and upper-tail dependent, whereas the Gaussian copula shows no tail preferences.

For the bivariate case, there is a rich number of practical and well-studied copulas. However, for higher dimensions, the application of copulas is challenging. Although multivariate Gaussian and multivariate t-copulas are widely used (Mashal and Zeevi, 2002), they cannot fully capture different dependence structures for different pairs of variables; all pairwise relationships are forced to follow the same copula. Several generalisations of bivariate copulas to higher-dimensional Archimedean copulas have been put forward (Savu and Trede, 2009), but they impose undesirable constraints on the parameter estimates (Martey and Attoh-Okine, 2019).

4.2.4 Vine copulas

Pioneered by Joe (1996) and further developed by Bedford and Cooke (2002) and Aas et al. (2009), the vine copula overcomes such shortcomings. It is a more natural and flexible way of formulating a high-dimensional copula based on a series of bivariate copulas, or so-called pair-copulas. This Pair-Copula Construction (PCC) methodology decomposes a multivariate copula density, and thus a multivariate probability density, into a product of (conditional) bivariate copulas, where all pair-copulas can be modelled independently from each other. It follows that a suitable bivariate copula can be freely chosen from a broad set of options to model the different dependence characteristics (including independence) of each variable pair, providing much greater flexibility in modelling dependence for high-dimensional data. Through a financial application, Aas et al. (2009) compared a vine copula containing Student copulas for pairs of stocks with the four-dimensional Student copula. A likelihood ratio test favoured the pair-copula construction method over the four-dimensional Student copula. Also, they found that the latter could lead to a large trading portfolio loss due to its underestimation of tail dependence. In a structural credit risk model setting, similar conclusions were drawn by Geidosch and Fischer (2016), who demonstrated that the estimation of economic capital for credit portfolios is more accurate when vines are employed rather than conventional copulas to model dependencies between latent asset values.

In conclusion, the vine copula provides considerable flexibility in modelling multivariate distributions by: (1) isolating the marginal and dependence formulations; and (2) matching the dependence structure of each respective variable pair with the most appropriate bivariate copula. However, this flexibility comes at a cost, in that the pair-copula construction has no unique representation due to the substantial number

of possible vine structures. To help organise them, Bedford and Cooke (2002) have introduced the regular vine (R-vine) and illustrated each possible decomposition of the bivariate copula density as a graphical tree. Two popular subclasses of R-vine were subsequently developed: the Canonical C-vine and the Drawable D-vine (Aas et al., 2009). They have been applied actively in financial and insurance risk management; see, for example, Nikoloulopoulos et al. (2012) and Schirmacher and Schirmacher (2008).

4.2.5 Vine copula-based quantile regression

This paper adopts the D-vine copula-based quantile regression model, proposed by Kraus and Czado (2017) and Schallhorn et al. (2017), to analyse the conditional EAD quantiles, taking into account the complex high-dimensional interrelationships among EAD and its predictors. The correlations between the predictors themselves are also considered, which are not commonly analysed in the literature. This interval estimation approach addresses several drawbacks of classical quantile regression including quantile crossing and multicollinearity problems. It also does not impose a restrictive linearity assumption on the shape of conditional quantiles and allows for the separation of marginal and dependence modelling. The model is fitted using a novel algorithm developed by Kraus and Czado (2017). This sequentially fits the D-vine structure with the aim of maximising a conditional likelihood, resulting in automatic variable selections. Due to the model construction, the conditional quantiles can be extracted easily from a series of estimated pair-copulas and do not cross each other. To the best of the authors' knowledge, this paper is the first to propose the vine copula-based quantile regression framework in any credit risk setting.

4.2.6 Research contributions

To summarise, the contributions of our research are that: (1) it is the first study to provide interval estimates and quantile predictions for EAD based on classical linear quantile regression and a state-of-the-art alternative — vine copula-based quantile regression; (2) we show that, on a large real-world credit card dataset, the latter model with non-parametric copulas performs better than the OLS linear model in terms of the point and interval estimates, conditional quantiles, and the distributions that they produce; (3) our results provide new insights into the predictor effects at different quantile levels of the EAD distribution, rather than on the mean level only; (4) we introduce the idea that complex multi-dimensional dependencies among loan-level variables can be effectively modelled using vine copulas, which has further potential applications to other consumer credit risk parameters such as PD and LGD.

4.3 Data and variables

The data from which our sample is extracted consists of monthly account-level data for the consumer credit cards of a large Hong Kong bank, recorded between January 2002 and May 2007. EAD is measured as the outstanding balance at default, excluding any subsequent interests and additional fees. The default definition is that borrowers either: (1) missed or could not pay the agreed minimum payment for 90 consecutive days or more; (2) were declared bankrupt; or (3) the money they owed was charged off by the bank. Similarly to other work on EAD, we extract only the defaulted account data, to ensure that the predicted balance is conditional on default. To construct the sample, we use the standard yearly cohort method (Moral, 2006) and set the reference month to the 1^{st} November of each year. For each such yearly default cohort, we collect the values of the covariates a month prior to the reference month, namely in October, whereas the response value (EAD) is the observed balance in the subsequent month where the default occurs. Accounts that lack sufficient monthly records to calculate the explanatory variables are omitted.

Table 4.1 lists the explanatory variables; all of these were shown to have a significant relationship with EAD according to previous literature; see e.g. Tong et al. (2016). After removing a small number of missing value cases, the total number of accounts used in the analysis is 63,476. We randomly divide this dataset into an in-sample training (80%) and out-of-sample test (20%) set. Note that there is no validation set because the process of selecting non-parametric distributions and input variables will be performed automatically by the fitting algorithm applied in the proposed model. Following Van Gestel et al. (2006), outliers are handled by winsorisation, by truncating outliers at $m \pm 3s$, where m is the median, $s = \frac{\text{IQR}}{2\times 0.6745}$, and IQR is the interquartile range.

Variable	Notation	Explanation
Age of account	age	Months since account has been opened.
Limit	l	Credit limit, i.e. maximum amount that can be drawn from card.
Balance	b	Current amount drawn.
Behavioural score	bsco	Internal score capturing current credit quality of account.
Average paid percentage past 9 months	paid.per9	Paid percentage is the percentage of last month's balance paid by the borrower, i.e. paid amount/balance.
Credit utilisation	cu	Percentage of the limit drawn by borrower, i.e. balance/limit.
Full payment percentage	full.pay.per	Percentage of account's months on book in which borrower has paid balance in full, i.e. number of full payments / age of account.

TABLE 4.1: List of available explanatory variables.

Figure 4.1 presents the pairwise scatter plots of all variables (extracted from a random sample of the full dataset for clearer visualisation), with histograms shown on the main diagonal. This exploratory analysis points to various non-normal marginal distributions and the presence of heteroscedasticity (see e.g. limit versus balance), which quantile regression should be capable of handling. Moreover, several of the bivariate relationships between predictors and EAD appear to be non-linear, and there are pronounced correlations between the predictors themselves, with some apparent asymmetric and tail dependencies that vary from one pair to another. This supports the application of copulas and suggests potential benefits to applying the proposed combined approach of vine copula-based quantile regression.

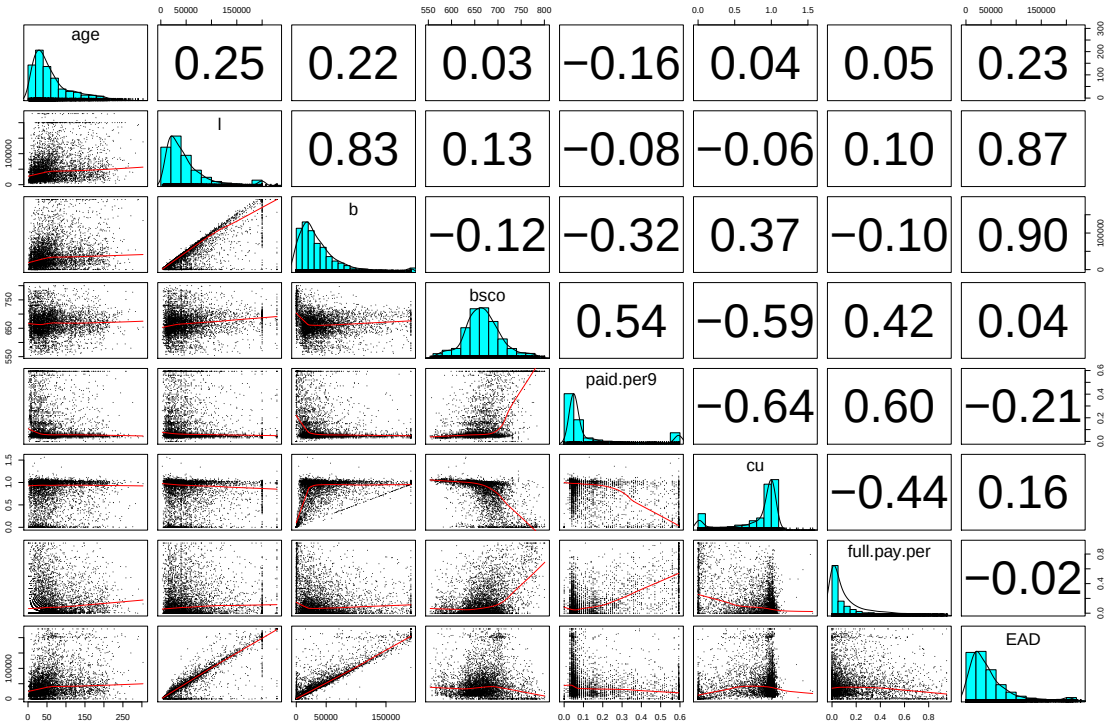


FIGURE 4.1: Pairwise scatter plots with histogram extracted from a partial set of EAD data; pairwise correlations are shown in the section above the main diagonal.

4.4 Vine copulas

A brief description of vine copulas is provided in this section. The joint multivariate distribution F of $\mathbf{X} = (X_1, \ldots, X_p)$ can be constructed by utilising Skalr's theorem (Sklar, 1959): for the marginal univariate distributions $F_1, \ldots, F_p$, there exists a copula function $C\colon [0,1]^p \to [0,1]$ such that $F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$. The copula approach allows the variable margins $F_j, j = 1, \ldots, p$, to be chosen from arbitrary distributions and modelled independently from their dependence structure (reflected by a chosen copula C). The copula C is unique when the corresponding cumulative

marginal distribution functions in $\mathbf{X}$ are continuous. Under further regularity conditions, the joint multivariate density of $\mathbf{X}$ can be written as:

$$f(x_1 \ldots, x_p) = c(F_1(x_1), \ldots, F_p(x_p)) \cdot \prod_{i=1}^{p} f_i(x_i), \tag{4.1}$$

where $f_1, \ldots, f_p$ are the marginal densities, and $c(u_1, \ldots, u_p) = \frac{\partial^p}{\partial_{u_1} \cdots \partial_{u_p}} C(u_1, \ldots, u_p)$ is the copula density. The p-dimensional $c(u_1, \ldots, u_p)$ can be decomposed into a product of $\frac{p(p-1)}{2}$ (conditional) bivariate copula densities, or so-called pair-copulas (Bedford and Cooke, 2001). Following Aas et al. (2009), a D-vine Pair-Copula Construction (PCC) with order $X_1 \to X_2 \to \ldots \to X_p$ of the joint density f can be written as:

$$f(x_1 \ldots, x_p) = \prod_{k=1}^{p} f_k(x_k) \prod_{i=1}^{p-1} \prod_{j=i+1}^{p} c_{ij|i+1,\ldots,j-1}(F_{i|i+1,\ldots,j-1}(x_i|x_{i+1}, \ldots, x_{j-1}), \\ F_{j|i+1,\ldots,j-1}(x_j|x_{i+1}, \ldots, x_{j-1})|x_{i+1}, \ldots, x_{j-1}), \tag{4.2}$$

where for a set $D \subset \{1, \ldots, p\}$ and $i, j \in \{1, \ldots, p\} \backslash D$, given $X_D = x_D$, $c_{ij|D}(\cdot, \cdot|x_D)$ is the (conditional) bivariate copula density associated with the conditional distributions $F_{i|D}(x_i|X_D = x_D)$ and $F_{j|D}(x_j|X_D = x_D)$. A common simplifying assumption of the pair-copulas is made here that $c_{ij|D}$ does not depend on the conditioning vector X_D, i.e. $c_{ij|D}(\cdot, \cdot|x_D) = c_{ij|D}(\cdot, \cdot)$. For more explanations, see Stöber et al. (2013). If all marginal distributions are uniformly distributed, the PCC is called a D-vine copula. We exemplify a four-dimensional D-vine copula with order $X_1 \to X_2 \to X_3 \to X_4$:

$$\begin{aligned} f(x_1, x_2, x_3, x_4) =& f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \cdot f_4(x_4) \cdot c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \cdot \\ & c_{34}(F_3(x_3), F_4(x_4)) \cdot c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \cdot \\ & c_{24|3}(F_{2|3}(x_2|x_3), F_{4|3}(x_4|x_3)) \cdot c_{14|23}(F_{1|23}(x_1|x_2, x_3), F_{4|23}(x_4|x_2, x_3)). \end{aligned} \tag{4.3}$$

This example clearly depicts an advantage of vine copulas, that is, each pair-copula can be chosen independently from each other to match the dependency pattern between the associated variable pair seen in the data. The first commonly-used class of bivariate copulas are parametric copulas, which comprise two main families: the elliptical copulas (e.g. Student-t and Gaussian) and the Archimedean copulas (e.g. Frank, Gumbel, and Joe). However, parametric copulas bear the risk of being wrongly specified and are likely to be inefficient when handling data-specific dependence structures such as non-monotonic relationships (Dette et al., 2014). As a remedy, the second class of non-parametric copulas has been proposed. Penalised and non-penalised Bernstein polynomials were utilised by Kauermann and Schellhase (2013) and Scheffer and Weiß (2016), respectively, whilst Nagler and Czado (2016) applied kernel estimators. We adopt the kernel weighted local likelihood technique, based on a common transformation trick introduced in Nagler et al. (2017), to estimate non-parametric bivariate copulas, because it has been proved (Nagler et al., 2017) to

perform best among the aforementioned methods if there is a strong tail dependence between the variables (which is our expected scenario for the EAD dataset).

Since the variables of interest, X_j, can be assigned exchangeably, the vine copula structures are not unique and could be represented in an abundance of combinations, especially for high-dimensional data. To help organise them, Bedford and Cooke (2002) depicted vine copulas through a nested sequence of trees known as dependence trees. Figure 4.2 displays a four-dimensional D-vine structure from Equation (4.3). The marginal densities f_1, f_2, f_3, f_4 are the nodes in the first tree T_1, whereas each edge, connected by the nodes, represents a pair-copula. The nodes for a tree T_{j+1} are then formed by the edges of a lower tree $T_j, j = 1, \ldots, p-2$, and the construction of nodes and edges for the subsequent trees is sequentially performed until the last tree T_{p-1}. Hence, the D-vine tree is useful for decomposing the multivariate copula into a product of bivariate copulas because the initial tree, T_1, can determine the entire structure.

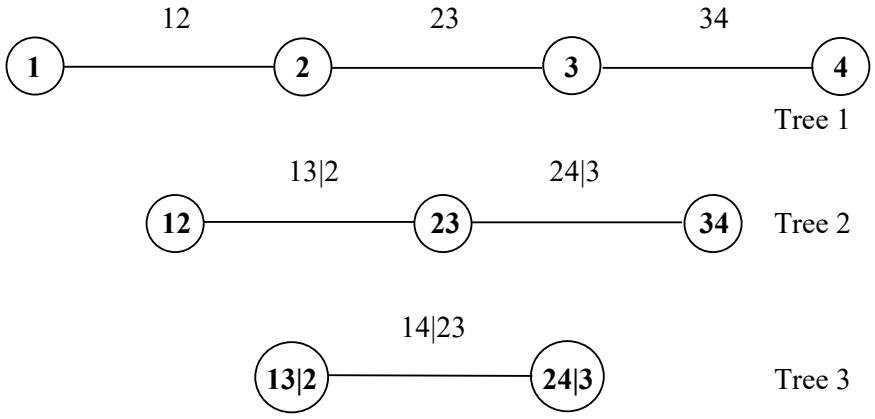


FIGURE 4.2: A four-dimensional D-vine with order $X_1 \to X_2 \to X_3 \to X_4$; each edge represents a pair-copula.

The conditional distributions $F_{i|D}(x_i|x_D)$ in Equation (4.2) can be estimated recursively based on pair-copulas from the respective lower trees, as follows:

$$F_{i|D}(x_i|x_D) = h_{il|D_{-l}}(F_{i|D_{-l}}(x_i|x_{D_{-l}}), F_{l|D_{-l}}(x_l|x_{D_{-l}})), \tag{4.4}$$

where $l \in D$ and $D_{-l} := D \backslash \{l\}$, and for $i, j \notin D$ and $i < j$, the h-functions associated with the (conditional) bivariate copula function $C_{ij|D}$ are defined as $h_{ij|D}(u, v) = \frac{\partial C_{ij|D}(u,v)}{\partial v}$ and $h_{ji|D}(u, v) = \frac{\partial C_{ij|D}(u,v)}{\partial u}$. For example, the first component $F_{1|23}(x_1|x_2, x_3)$ of $c_{14|23}$ from Tree 3 (in Figure 4.2) can be evaluated via the h-functions

related to $C_{13|2}$, C_{12}, and C_{23} from the first two trees:

$$\begin{aligned} F_{1|23}(x_1|x_2, x_3) &= h_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \\ &= h_{13|2}(h_{12}(F_1(x_1), F_2(x_2)), h_{32}(F_3(x_3), F_2(x_2))). \end{aligned}$$

Hence, Equation (4.4) allows us to estimate the joint multivariate density, $f(x_1 \ldots, x_p)$, in Equation (4.2), from the marginal univariate distributions, $F_1, \ldots, F_p$, and pair-copulas, C_{ij}.

4.5 Statistical models

In this section, we explain how to predict the conditional quantile of the response (Exposure At Default), $Y \sim F_Y$, given the outcome of a set of p continuous covariates, $X_j \sim F_j, j = 1, \ldots, p$, from either the proposed D-vine copula-based quantile regression model or a classical linear quantile regression model. An OLS linear regression is also specified, which will serve as a benchmark for a subsequent performance comparison.

4.5.1 D-vine copula-based quantile regression

In the D-vine copula-based quantile regression model (henceforth referred to DVQR), the conditional α quantile, for $\alpha \in (0, 1)$, is calculated as:

$$q_\alpha(x_1, \ldots, x_p) \coloneqq F^{-1}_{Y|X_1,\ldots,X_p}(\alpha|x_1, \ldots, x_p), \tag{4.5}$$

where F is the multivariate joint distribution of $Y, X_1, \ldots, X_p$ established from a D-vine copula. By using Skalr's theorem and the probability integral transform (PIT), $V \coloneqq F_Y(Y)$ and $U_j \coloneqq F_j(X_j)$ with corresponding PIT values $v \coloneqq F_Y(y)$ and $u_j \coloneqq F_j(x_j)$, we obtain:

$$\begin{aligned} F_{Y|X_1,\ldots,X_p}(y|x_1, \ldots, x_p) =& P(Y \leq y|X_1 = x_1, \ldots, X_p = x_p) \\ =& P(F_Y(y) \leq v|F_1(X_1) = u_1, \ldots, F_p(X_p) = u_p) \\ =& C_{V|U_1,\ldots,U_p}(v|u_1, \ldots, u_p). \end{aligned}$$

That is, $C_{V|U_1,\ldots,U_p}$ is the conditional distribution of V given $(U_1, \ldots, U_p)$ associated with the conditional distribution function of Y given $(X_1, \ldots, X_p)$. Thus, Equation (4.5) can be expressed as follows:

$$\begin{aligned} q_\alpha(x_1, \ldots, x_p) &= F_Y^{-1}(C^{-1}_{V|U_1,\ldots,U_p}(\alpha|u_1, \ldots, u_p)) \\ &= F_Y^{-1}(C^{-1}_{V|U_1,\ldots,U_p}(\alpha|F_1(x_1), \ldots, F_p(x_p))). \end{aligned} \tag{4.6}$$

Hence, the conditional quantile can be derived by estimating the univariate distributions F_Y and F_j and the $(p+1)$-dimensional copula $C_{V,U_1,\dots,U_p}$. This shows that DVQR permits us to separately model the margins and their dependencies, and does not make any restrictive assumptions on the shape of conditional quantiles. Note that the closed form of the conditional quantile can be expressed only in a purely continuous setting. In contrast, if there are discrete variables, we need to refer to Schallhorn et al. (2017) and compute the conditional quantile by numerically inverting the conditional distribution function.

The conditional quantile, shown in Equation (4.6), can be extracted analytically by applying the recursion in Equation (4.4) and expressing $C_{V,U_1,\dots,U_p}$ in terms of nested h-functions. A four-dimensional example is provided below.

$$
\begin{aligned}
&C_{V|U_1,U_2,U_3}(v|u_1,u_2,u_3)\\
&= h_{V,U_3|U1,U2}(C_{V|U_1,U_2}(v|u_1,u_2), C_{U_3|U_1,U_2}(u_3|u_1,u_2))\\
&= h_{V,U_3|U1,U2}(h_{V,U_2|U_1}(C_{V|U_1}(v|u_1), C_{U_2|U_1}(u_2|u_1)), h_{U_3,U_1|U_2}(C_{U_3|U_2}(u_3|u_2), C_{U_1|U_2}(u_1|u_2)))\\
&= h_{V,U_3|U1,U2}(h_{V,U_2|U_1}(h_{V,U_1}(v,u_1), h_{U_2,U_1}(u_2,u_1)), h_{U_3,U_1|U_2}(h_{U_3,U_2}(u_3,u_2), h_{U_1,U_2}(u_1,u_2))),
\end{aligned}
$$

the inverted function of which is

$$
\begin{aligned}
&C^{-1}_{V|U_1,U_2,U_3}(\alpha|u_1,u_2,u_3)\\
&= h^{-1}_{V,U_1}[h^{-1}_{V,U_2|U_1}\{h^{-1}_{V,U_3|U_1,U_2}(\alpha, h_{U_3,U_1|U_2}(h_{U_3,U_2}(u_3,u_2), h_{U_1,U_2}(u_1,u_2))), h_{U_2,U_1}(u_2,u_1)\}, u_1].
\end{aligned}
$$

Since $C^{-1}_{V|U_1,\dots,U_p}(\alpha|u_1,\dots,u_p)$ is monotonically increasing with α, the problem of different α quantile functions crossing each other is naturally eliminated (Kraus and Czado, 2017).

We consider two submodels: parametric DVQR (P-DVQR), in which bivariate copulas are chosen exclusively from parametric families, and non-parametric DVQR (NP-DVQR), where bivariate copulas are estimated non-parametrically. The former includes Gaussian, Student-t, Clayton, Gumbel, Joe, Frank, Clayton-Gumbel, Joe-Gumbel, Joe-Clayton, Joe-Frank copulas, and their rotations (Nelsen, 2006). The estimation of the variable distributions F_Y and F_j and the copula $C_{V,U_1,\dots,U_p}$ are performed in two steps, using a recent computational method for the DVQR proposed by Kraus and Czado (2017) and implemented in the R package vinereg (Nagler and Kraus, 2019). First, the marginal distributions, F_Y and F_j, are estimated non-parametrically by a kernel smoothing method (Parzen, 1962). Given a sample $(x^{(1)},\dots,x^{(n)}) \in \mathbb{R}^n$, where n is the number of observations, the estimator is $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x-x^{(i)}}{h})$, where $K(x) := \int_{-\infty}^{x} k(t)\,dt$ with $k(\cdot)$ being a symmetric probability density function and $h > 0$ a bandwidth parameter developed in Duong (2016). More specifically, the bandwidth is defined as:

$$h = \underset{h}{\operatorname{argmin}} \int_{-\infty}^{\infty} E(F(x) - \hat{F}(x))^2 dx,$$

i.e. it minimises the mean integrated squared error of the estimated kernel distribution. This bandwidth selection approach provides a desirable property as it optimally finds a trade-off between oversmoothing (more smooth fitted curve with correspondingly less accuracy) and undersmoothing. Subsequently, the estimated $\hat{F}_Y$ and $\hat{F}_j$ are used to transform the data from their original scale to pseudo copula data in $[0,1]$ scale: $\hat{v}^{(i)} := \hat{F}_Y(y^{(i)})$ and $\hat{u}_j^{(i)} := \hat{F}_j(x_j^{(i)}), j = 1, \ldots, p, i = 1, \ldots, n$.

In the second step, the multivariate copula $C_{V,U_1,\ldots,U_p}$ is fitted by a D-vine copula with the copula data generated from the previous step. Two stages are involved: establishing the dependence (vine) structure and drawing statistical inferences on pair-copulas. First, the vine is constructed by fixing the response V at the initial node in the first tree and choosing the order of other covariate variables U_j with the objective of maximising the predictive strength of the model. The order (from high to low) of the explanatory power of a covariate is therefore reflected by its position in the first tree (from left to right). An algorithm similar to a forward stepwise method is employed. Hence, variable selection is accomplished automatically, by sequentially adding the most influential covariate that improves the model's fit, measured by the conditional log-likelihood for the response given the set of covariates, i.e. $\sum_{i=1}^{n} \log c_{V|U_1,\ldots,U_p}(v^{(i)}|u_1^{(i)}, \ldots, u_p^{(i)})$, where $c_{V|U_1,\ldots,U_p}$ is the copula density associated with $C_{V|U_1,\ldots,U_p}$. This process continues until no additional improvement can be obtained. Second, a bivariate copula selection is performed based on the Akaike Information Criterion (AIC). Denote the ordering vector $(l_1, \ldots, l_p)^T$ a permutation of $(1, \ldots, p)^T$ demonstrating the covariate's order in the vine tree. When a new covariate $U_{l_k}, k = 2, 3, \ldots, p$, is being added to the current D-vine with order $V \to U_{l_1} \to \ldots \to U_{l_{k-1}}$, the AIC-optimal pair-copulas and their parameters (Genest and Favre, 2007) are selected from different choices of bivariate copulas. This process determines the pair-copulas between the response and the new covariate, $\hat{C}_{V,U_{l_k}|U_{l_1},\ldots,U_{l_{k-1}}}$, as well as those among the existing covariates and the new covariate, $\hat{C}_{U_{l_1},U_{l_k}|U_{l_2},\ldots,U_{l_{k-1}}}, \hat{C}_{U_{l_2},U_{l_k}|U_{l_3},\ldots,U_{l_{k-1}}}, \ldots, \hat{C}_{U_{l_{k-1}},U_{l_k}}$. To tackle a wide range of dependencies, we consider the Gaussian (N), Student-t (t), Clayton (C), Gumbel (G), Joe (J), Frank (F), Clayton-Gumbel (BB1), Joe-Gumbel (BB6), Joe-Clayton (BB7), Joe-Frank (BB8) copulas, and their rotations (Nelsen, 2006), as potential parametric choices; in addition, we consider the independence copula and a transformation kernel technique for the non-parametric choices (Nagler et al., 2017). These estimated pair-copulas are the basis of h-functions used to calculate $\hat{C}_{V,U_1,\ldots,U_p}$ and, hence, the conditional quantile as shown in Equation (4.6). Figure 4.3 summarises the procedures in the estimation process.

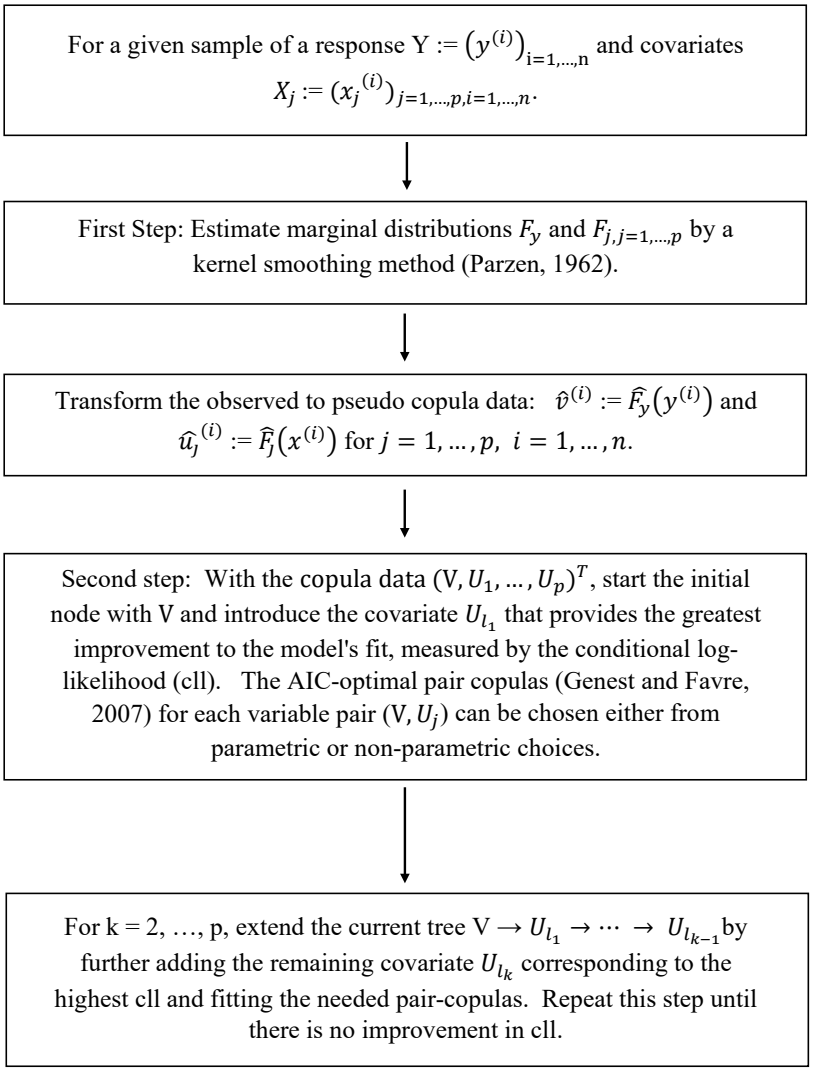


FIGURE 4.3: Scheme of the DVQR estimation process.

Therefore, the proposed estimation process results in a parsimonious flexible model, avoids multicollinearity problems, and removes the need for variable transformations due to the relaxed assumptions on how the covariates influence the response and the flexible distribution class for marginals. The computational time is reliably fast. For example, the elapsed time for fitting the NP-DVQR model with 50,750 dataset is six minutes on the high performance IRIDIS 5 compute cluster with dual 2.0 GHz Intel Skylake processors and 192 GB of DDR4 memory. For extensive details, see Kraus and Czado (2017).

4.5.2 Linear quantile regression

The predicted conditional quantile derived from a linear quantile regression (referred to as LQR) (Koenker and Bassett, 1978) is assumed to be linear in the predictors, i.e.

$$\hat{q}_\alpha(x_1^{(i)}, \ldots, x_p^{(i)}) := \hat{\beta}_0(\alpha) + \sum_{j=1}^{p} \hat{\beta}_j(\alpha) x_j^{(i)}, \tag{4.7}$$

where $i = 1, \ldots, n$. It allows each quantile to be modelled individually by separate regressions. The unknown parameters $\hat{\boldsymbol{\beta}}(\alpha) \in \mathbb{R}^{p+1}$ are estimated with the minimisation problem

$$\min_{\boldsymbol{\beta}(\alpha) \in \mathbb{R}^{p+1}} \rho_\alpha(y^{(i)} - (\beta_0(\alpha) + \sum_{j=1}^{p} \beta_j(\alpha) x_j^{(i)})),$$

where $\rho_\alpha(u) = u(\alpha - \mathbb{I}(u < 0))$ and $\mathbb{I}$ is an indicator function. In contrast to a symmetrically quadratic loss function used in the OLS, here, residuals are weighted by an asymmetric loss function ρ_α. For upper quantile levels $\alpha \in (0.5, 1)$, positive residuals, or equivalently underestimations, are subjected to heavier loss by the weight $\alpha \in (0.5, 1)$ than negative residuals (overestimations) by the weight $1 - \alpha$. This results in an unbiased, consistent, and asymptotically normally distributed estimator for the α quantile regression (Krüger and Rösch, 2017).

4.5.3 Linear regression

We introduce the OLS linear regression model (referred to as OLS) as a benchmark model with the formulation

$$\hat{Y} | X_1^{(i)}, \ldots, X_p^{(i)} := \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j^{(i)} + \epsilon^{(i)}, \tag{4.8}$$

where the errors $\epsilon^{(i)}$ are assumed to be independent of each other with mean zero and equally constant variance σ^2. The unknown parameters $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ are estimated with the minimisation problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} (y^{(i)} - (\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)}))^2.$$

We calculate the predicted conditional quantile by fitting a normal distribution for estimated errors. Hence, the conditional $\hat{Y} | X_1^{(i)}, \ldots, X_p^{(i)}$ is normally distributed with mean $\mu^{(i)} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j^{(i)}$ and variance σ^2, and $\hat{q}_\alpha(x_1^{(i)}, \ldots, x_p^{(i)}) := N^{-1}(\alpha | \mu^{(i)}, \sigma^2)$, where N^{-1} is the inverse normal CDF. Note that the choice of benchmarking the OLS model for quantile regression is because it is a standard method commonly applied in

EAD modelling literature. Moreover, the implementation of OLS allows us to inspect the changes in predictor effects when different levels of EAD quantiles are focused rather than its mean level.

4.6 Analyses and results

4.6.1 Parameter estimates

Variable	$\tau = 0.025$	$\tau = 0.50$	$\tau = 0.975$	OLS
(Intercept)	-720 (1059)	-14144** (879)	-101834** (5113)	-52840** (1799)
Age of account	-0.62 (0.87)	-4.72** (0.13)	40.68** (7.96)	-6.40** (1.64)
Limit	-0.09** (0.01)	0.35** (0.03)	1.58** (0.07)	0.32** (0.004)
Balance	1.01** (0.02)	0.71** (0.03)	-0.33** (0.07)	0.78** (0.005)
Behavioural score	3.61* (1.48)	24.27** (0.53)	145** (6.96)	91.18** (2.53)
Paid percentage	-1999** (572)	-177 (195)	10.83 (2660)	-7814** (685)
Credit utilisation	-1477** (377)	-1434* (785)	13802** (2161)	-5367** (443)
Full payment percentage	-2820** (677)	-35.34 (76.78)	1861 (2082)	-3469** (483)

TABLE 4.2: Parameter estimates of the linear quantile model for the 0.025, 0.50, and 0.975 quantiles. Standard errors are given in parentheses by kernel estimates. Significance level is indicated by $^{*}(5\%)$ and $^{**}(1\%)$. The last column represents the OLS estimates.

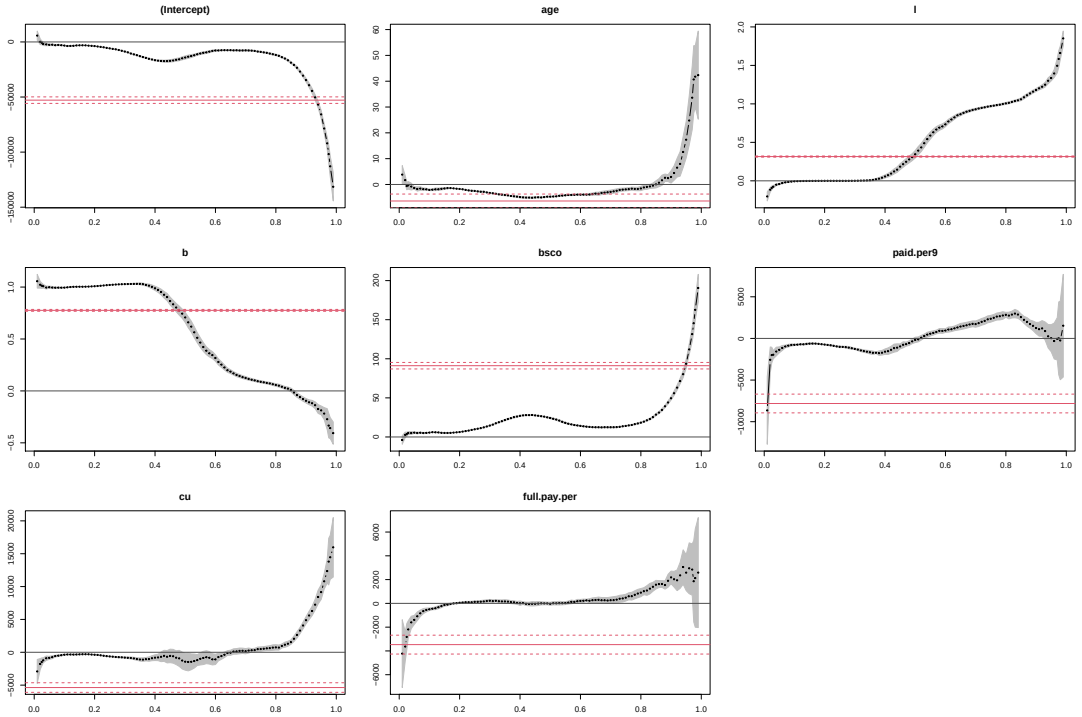


FIGURE 4.4: Parameter estimates with 95% confidence intervals of the linear (red) and linear quantile (grey) models with respective quantile levels on x-axis. Standard errors are estimated by kernel estimates.

In Table 4.2, we compare the parameter estimates of the linear quantile model (LQR) for the 0.025, 0.50, and 0.975 quantiles, with the coefficients of the OLS linear model (OLS). As observed from the table, the significance and level of the LQR estimates strongly depend on the associated predictive quantile. For example, the effects of paid percentage and full payment percentage are not statistically different from zero at the 0.5 and 0.975 quantiles, whereas, at the opposite end of the distribution, the EADs at the 0.025 quantile are not significantly related to the account's age. The full plots for all quantiles are provided in Figure 4.4. As depicted, OLS yields a single set of parameter estimates (for the conditional EAD mean), whereas LQR produces estimates that depend on the quantile being considered. For instance, the effect of the credit limit on the EAD appears negligible for the lower quantiles but very strong for the upper ones, which is intuitive as the limit acts as an upper ceiling. Some predictors also exhibit opposite effects; for example, utilisation rate has a positive effect on the right EAD tail whilst its impact on the left tail is negative, suggesting that greater card activity may widen the EAD distribution. Interestingly, most predictors, namely limit, balance, account's age, rating score, and credit utilisation, influence EAD more strongly at the upper quantiles, implying that they should feature more prominently when a more conservative EAD risk estimate (such as the 99% value-at-risk), as opposed to a point estimate for the conditional mean, is required. In contrast, OLS is unable to capture non-constant variable effects, leading to substantially different (and possibly distorted) parameter estimates compared to the LQR estimates.

4.6.2 Vine copula dependence structure

In this subsection, we analyse the selection of vine structure, as well as a set of pair-copulas and their respective estimated parameters, for the D-vine copula-based quantile regression models. An algorithm similar to a forward variable selection is used to determine the order of the first tree (and thus the complete structure) in D-vine, and the best fitting pair-copula for each variable pair is identified using the AIC criterion.

Figure 4.5 exhibits the estimated D-vine with parametric copulas (P-DVQR), where each row represents a tree and its respective edges, with the first tree located at the bottom. The chosen AIC-optimal pair-copulas result in the presented contour plots, reflecting the joint PDF of the variable pair; their maximum likelihood estimates and Kendall's tau are shown in Table 4.3. The bottom row of Figure 4.5 shows all variables ordered by their explanatory power, the leftmost (rightmost) variable being the strongest (weakest) predictor, respectively. Balance thus has the strongest effect on EAD, followed by limit, rating score, utilisation rate, full payment percentage, paid percentage, and account's age. The selection of the Student-t copula with a high Kendall's Tau of 0.75 (see top row of Table 4.3) suggests that the underlying

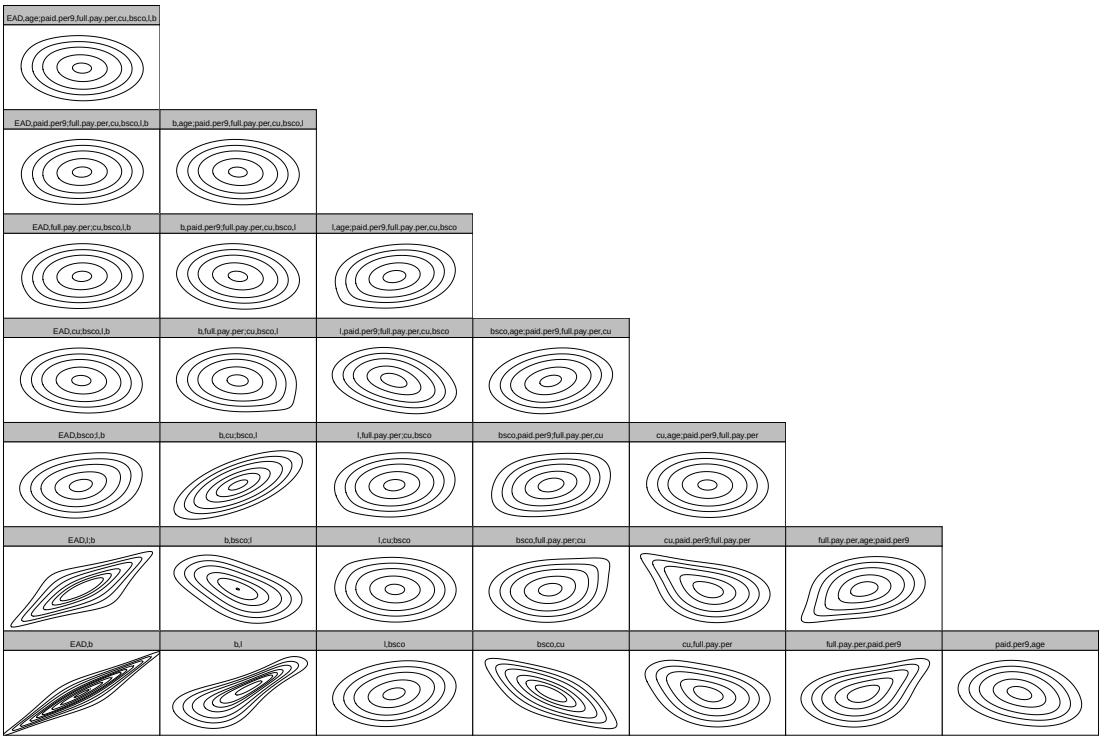


FIGURE 4.5: Estimated D-vine with parametric copulas and contour plots displaying the joint PDF of variable pairs with the first component on x-axis and the second on y-axis. The order of the D-vine is EAD → b → l → bsco → cu → full.pay.per → paid.per9 → age.

dependence between EAD and balance is strongly positive and symmetric, exhibiting both upper and lower tail dependence. That is, balance at default and current balance are expected to move in the same direction, especially in the tails of their distributions. Similarly, the conditional dependence between EAD and limit given balance (see the 13|4 variable pair entry of the second tree in Table 4.3) is also captured by the t copula but with a weaker rank correlation of 0.51. Rating score shows a decent correlation with EAD, with a mild positive upper tail dependence implied by the Joe-Frank (BB8) copula (see the 15|34 entry of the third tree in Table 4.3). The dependencies between EAD and the other covariates are relatively weak. Several strongly related variable pairs are also found among the explanatory variables themselves. Similarly to EAD and limit earlier, balance (now prior to default) and limit are highly correlated at high values but only mildly correlated elsewhere (see the second plot at the bottom row of Figure 4.5). A similar dependence pattern is seen for full payment percentage versus paid percentage (see the sixth plot at the bottom row of Figure 4.5), albeit to a lesser degree. Credit utilisation and credit score are also strongly related, exhibiting negative upper and lower tail dependencies (see the fourth plot at the bottom row of Figure 4.5); hence, higher (lower) card utilisation is indicative of a lower (higher) credit score, respectively. In summary, the majority of the selected pair-copulas are not symmetric and exhibit a range of different tail dependence patterns, which is not surprising for a financial dataset (see e.g. Kraus and Czado (2017)). Compared to a conventional correlation analysis, copulas thus provide deeper insights into the

Tree	Variable Pair	Copula family	Rotation	Parameter 1	Parameter 2	Tau
1	14	t	0	0.92	2.00	0.75
1	43	BB8	0	4.42	0.95	0.61
1	35	BB8	0	2.48	0.50	0.13
1	57	t	0	0.69	5.10	-0.48
1	78	BB8	90	1.78	0.92	-0.23
1	86	BB8	0	1.66	0.99	0.26
1	62	BB8	270	1.92	0.74	-0.16
2	13|4	t	0	0.72	2.18	0.51
2	45|3	BB8	90	4.23	0.64	-0.36
2	37|5	BB8	90	1.06	0.99	-0.03
2	58|7	BB8	0	1.20	1.00	0.11
2	76|8	BB7	90	1.55	0.01	-0.24
2	82|6	BB7	180	1.34	0.10	0.19
3	15|34	BB8	0	1.49	0.85	0.12
3	47|53	t	0	0.66	29.56	0.46
3	38|75	BB7	180	1.06	0.11	0.08
3	56|87	t	0	0.22	9.65	0.14
3	72|68	t	0	-0.03	50.00	-0.02
4	17|534	BB1	90	0.07	1.01	-0.04
4	48|753	Joe	270	1.1	-	-0.07
4	36|875	BB8	270	2.38	0.66	-0.19
4	52|687	BB8	0	3.31	0.35	0.12
5	18|7534	BB8	180	1.06	0.99	0.03
5	46|8753	BB8	270	2.14	0.39	-0.07
5	32|6875	BB1	180	0.08	1.09	0.12
6	16|87534	BB8	180	1.08	0.95	0.03
6	42|68753	BB8	270	1.34	0.59	-0.04
7	12|687534	BB8	90	1.09	0.91	-0.03

TABLE 4.3: Maximum likelihood estimates and Kendall's tau for AIC-optimal pair copulas. The variables are (1) EAD, (2) Age of account, (3) Limit, (4) Balance, (5) Behavioural score, (6) Average paid percentage past 9 months, (7) Credit utilisation, (8) Full payment percentage.

relationships between EAD and the other variables of interest.

Since parametric copulas could wrongly specify non-monotonic dependence structures (which are observed in our EAD dataset, see Figure 4.1), we extend the CGAMLSS analysis to also include non-parametric copulas. Figure 4.6 displays an estimated D-vine with non-parametric copulas (NP-DVQR). The D-vine order of NP-DVQR resembles that of P-DVQR with a slight difference in the order of utilisation rate and full payment percentage. None of the pair-copulas are modelled by the independence copula, which supports the existence of multicollinearity among our variables of interest. For the most part, the dependence structures of the estimated non-parametric pair-copulas are similar to their parametric counterparts. However, they reflect more realistic characteristics of the variables, and thus avoid misspecification. For instance, for the first two edges in the first tree, pair-copulas are

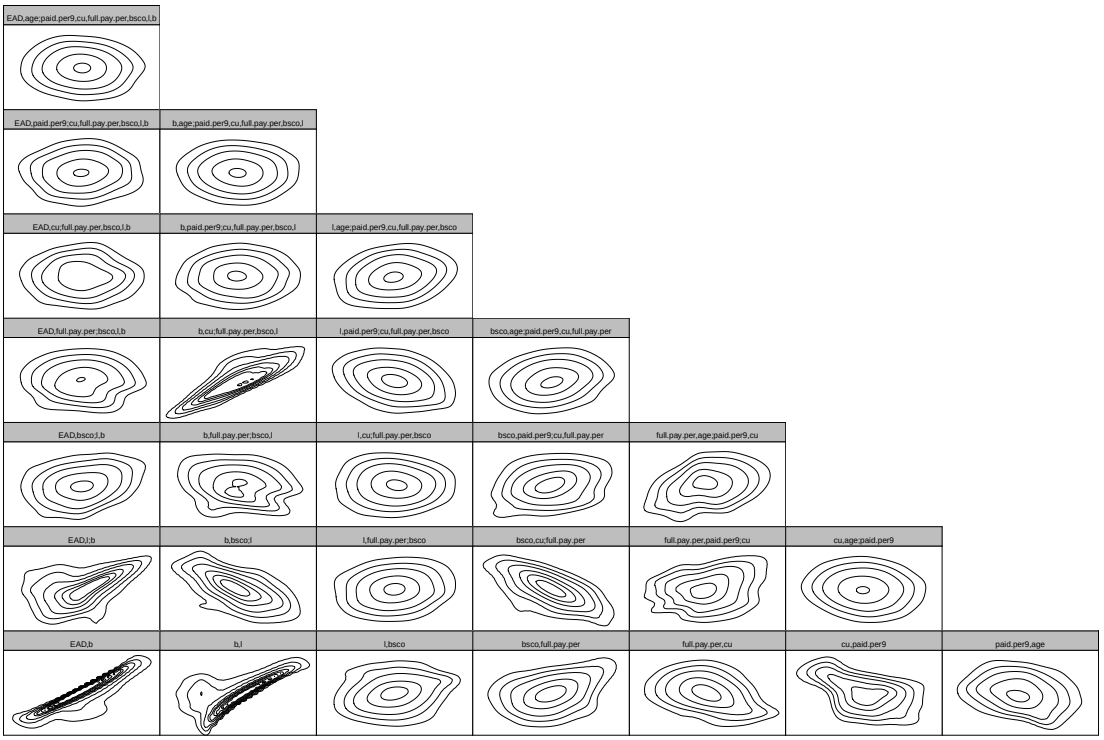


FIGURE 4.6: Estimated D-vine with non-parametric copulas and contour plots displaying the joint PDF of variable pairs with the first component on x-axis and the second on y-axis. The order of the D-vine is EAD → b → l → bsco → full.pay.per → cu → paid.per9 → age.

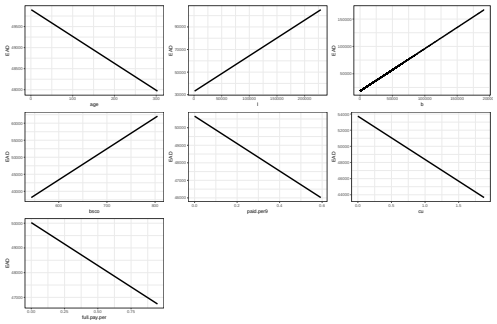
estimated so that EAD most of the time exceeds the balance prior to default (see the first edge, or lower-left plot, for EAD-b) and balance tends to be smaller than the limit (see the second edge, b-l), both of which are intuitive. In contrast, the P-DVQR results did not yet capture that exposure tends to increase in the race to default and that balance normally stays within the limit.

4.6.3 Effects of predictors

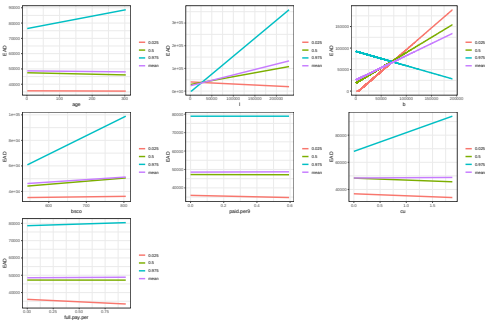
Figure 4.7 shows the partial effect plots for the different models, depicting how each predictor influences the response assuming that all other covariates are fixed at their respective mean levels. More specifically, they show the marginal effects on the conditional mean, $E(Y|X_1, \ldots, X_p)$, and on the 0.025, 0.5 and 0.975 conditional quantiles, $q_\alpha(x_1, \ldots, x_p)$, of EAD. The conditional mean for the quantile regression models is computed based on an average of a series of $\{1/11, 2/11, \ldots, 10/11\}$ quantiles.

In the OLS (top-left panel), the effects on EAD mean are, by definition, all linear; considering the scale on the y-axis, balance is the variable that has the largest effect. Next, LQR (top-right panel) is able to provide deeper insights into how these effects further vary depending on the EAD quantile of interest, showing that the impact of limit (l), credit score (bsco), and utilisation rate (cu) on the 0.975 quantile is much

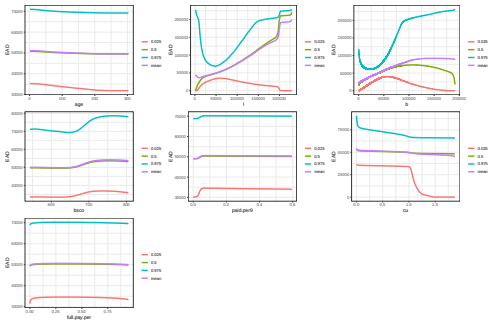
stronger than for the lower quantiles. Interestingly, the differing slopes in the LQR effect plots for limit and balance suggest that whereas limit is a key driver for the 0.975 quantile, balance is the more important driver for the 0.025 quantile. Also, 95% prediction intervals can be derived by contrasting the variable effect plots for the 0.025 and 0.975 quantiles. These suggest a much wider prediction interval and, hence, greater variability in EAD as the credit limit increases (again, keeping other variables constant). Conversely, paid percentage and full payment percentage, having roughly parallel effect plots, do not appear to impact the width of the prediction interval by much. The LQR estimates, however, are prone to quantile crossing. In the result plots



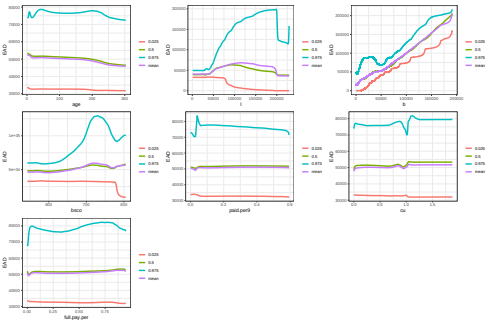
(a) Partial effect plots in OLS.



(b) Partial effect plots in LQR.



(c) Partial effect plots in P-DVQR.



(d) Partial effect plots in NP-DVQR.

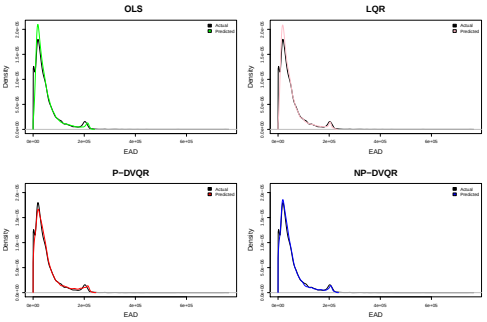
FIGURE 4.7: Partial effect plots of predictors on the conditional mean and 0.025, 0.5 and 0.975 conditional quantiles of EAD.

for limit and balance, the effect lines indeed cross each other, thus causing interpretation difficulties. For example, other things being equal, when balance exceeds 75,000, the top-right plot appears to suggest a lower EAD value at the 0.975 quantile than at the 0.025 quantile, which is clearly counter-intuitive. The D-vine copula models (DVQR), shown in the bottom panel of the figure, resolve this problem by computing quantiles from Equation (4.6) so that none of the effect lines cross each other. For example, in the effect plots for balance, the quantile order is now preserved. Another advantage of theirs is that the assumption of linearity is lifted, permitting conditional EAD quantiles to be non-linearly and non-monotonically related to the covariates. For example, some non-monotonicity is now observed with regards to the impact of the credit limit. Interestingly, the non-parametric model (NP-DVQR) is the

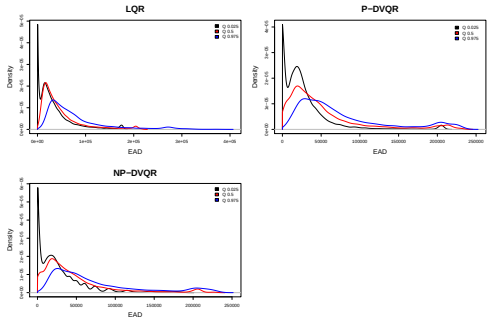
only to suggest a drop-off in EAD for the subgroup of accounts that were awarded the high credit limits at 200,000 by the bank.

4.6.4 EAD quantile distributions

Figure 4.8a compares the density plot for the actual EAD values with those for the point estimates (conditional EAD mean) produced by each model. Since EAD cannot take negative values, we fit the probability density function by zero-truncated kernel density estimation with a Gaussian kernel and weight $w(x) = \frac{1}{1-\Phi_{x,h}(0)}$, where h is the bandwidth and Φ is the cumulative distribution function of a Gaussian distribution with mean x and standard deviation h. The objective is to truncate the density on the negative side at zero and up-weight the data that are close to zero. We can see that the non-parametric DVQR provides the best fit to the empirical distribution, followed by the parametric DVQR model. Instead, OLS and LQR misspecify and overestimate EAD at the lower end. Hence, there is a positive gain to using the vine copula models. In the right panel, Figure 4.8b displays the density plots for three different conditional quantiles produced by LQR, P-DVQR and NP-DVQR. In line with expectation, the upper quantile (0.975) predictions all exhibit a heavy tail property. Among these, LQR produces the longest right tail, leading to the largest 97.5% value-at-risk for EAD.



(a) Density plots for the actual vs predicted EAD fitted by zero-truncated weighted kernel density estimates. Predicted EAD mean is used.



(b) Density plots of predicted EAD quantiles at 0.025, 0.5 and 0.975 quantile levels fitted by zero-truncated weighted kernel density estimates.

FIGURE 4.8: Density plots of predicted EAD.

4.6.5 Model performance

In order to evaluate how competitive the models are relative to each other, we conduct an out-of-sample predictive performance test containing n_{test} data points, where n_{test} is the sample size (20%) of the test set. We consider both the quality of the predicted EAD quantiles, as well as that of the point and interval estimates of EAD.

4.6.5.1 Accuracy of predicted quantiles

First, we inspect the predictive accuracy of the predicted conditional EAD quantiles at level $\alpha \in \{0.01, \ldots, 0.99\}$. Unlike the actual values observed in the test set, true regression quantiles remain unobserved. For that reason, Komunjer (2013) suggested the use of average α-weighted absolute error, WAE(α), defined as:

$$\text{WAE}(\alpha) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \rho_\alpha(y^{(i)} - \hat{q}_\alpha^{(i)}),$$

where $y^{(i)}$ is the actual value of EAD for the i-th observation in the test set, $\hat{q}_\alpha^{(i)} = \hat{q}_\alpha(x_1^{(i)}, \ldots, x_p^{(i)})$ is the predicted conditional α quantile, and $\rho_\alpha(u) = u(\alpha - \mathbb{I}(u < 0))$ is an asymmetric loss or check function. A lower WAE(α) denotes better performance. Second, as a counterpart to the coefficient of determination, the model fit is assessed by a goodness-of-fit measure, $R^1(\alpha)$, proposed by Koenker and Machado (1999):

$$R^1(\alpha) = 1 - \frac{\sum_{i=1}^{n_{train}} \rho_\alpha(y^{(i)} - \hat{q}_\alpha^{(i)})}{\sum_{i=1}^{n_{train}} \rho_\alpha(y^{(i)} - y_\alpha)},$$

where n_{train} is the sample size (80%) of the training set and y_α is the alpha quantile of all EAD values observed in the training set. The larger the $R^1(\alpha)$, the better the model fit. Haupt et al. (2011) stated that WAE(α) and $R^1(\alpha)$ seem to be a more natural way to evaluate the fit and predictive performance for L_1-norm based estimations such as quantile regressions rather than R^2 and the average absolute or squared errors.

Figure 4.9 thus depicts the performance of the conditional quantile predictions at $\alpha \in \{0.01, \ldots, 0.99\}$ for all four models. Where out-of-sample predictive accuracy is concerned (top panel), LQR and NP-DVQR produce the lowest weighted absolute errors and substantially outperform OLS for any quantile other than the median. Between the two vine copula models, the non-parametric one clearly outperforms the parametric one. A logical explanation for this lies in the presence of non-monotonic relationships between several pairs of variables in our dataset (see e.g. EAD and utilisation rate in Figure 4.1), which cannot be correctly modelled by a parametric copula (Dette et al., 2014). This misspecification appears to affect the model, making it perform even worse than the simple linear model at some of the quantiles. In addition, due to the fact that P-DVQR models the dependence of variable pairs through a set of parameter(s), it intrinsically imposes a particular dependence shape which could be different from the actual one. For example, the estimated t copula from the P-DVQR model (see the lower-left plot in Figure 4.5) exhibits a symmetrical pattern between EAD and balance which means EAD is allowed to be smaller than the balance, contrasting to the real observations. This parametric formulation might deteriorate the model performance. Although being relatively close, NP-DVQR

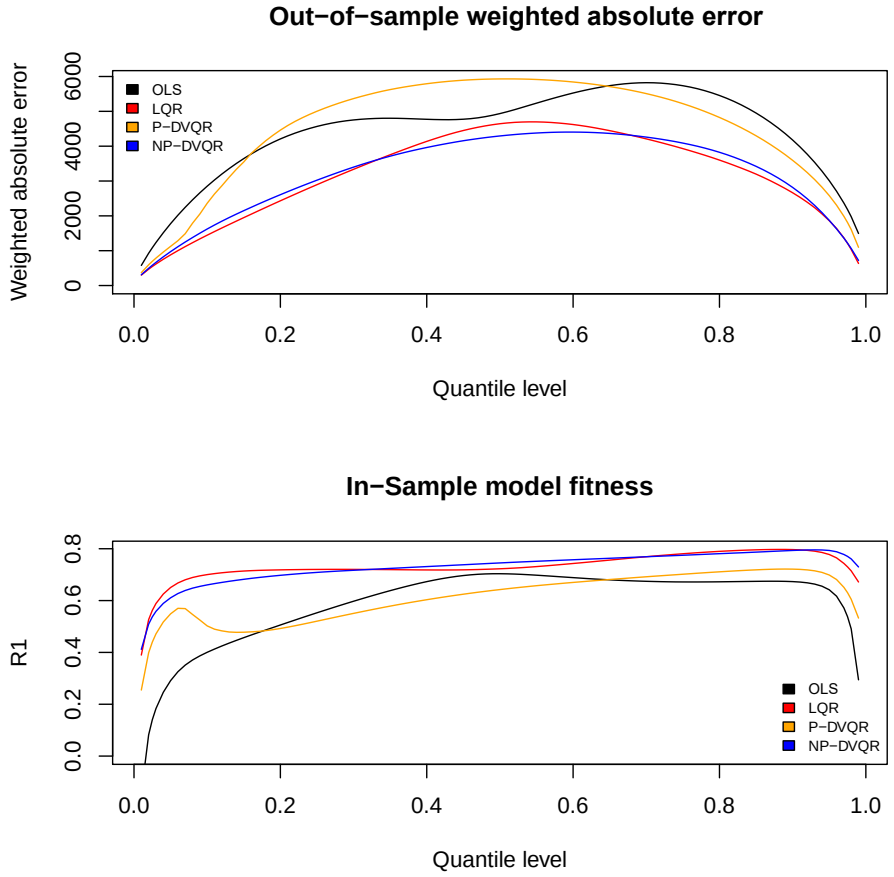


FIGURE 4.9: Performance measurements of the predicted conditional quantiles for OLS, LQR, P-DVQR and NP-DVQR: weighted absolute error (top) and model fitness (bottom).

performs better for the middle quantile predictions, whereas LQR is superior in the lower and upper tails. The model fitness results (bottom panel) lead to similar conclusions. In summary, in order to gain a better model for conditional EAD quantile estimation, one should apply a quantile regression method, specifically LQR or NP-DVQR, rather than a conventional linear model.

4.6.5.2 Quality of point and interval estimates

To evaluate the quality of the point estimates at the mean level, we use the mean absolute error (MAE) as the prediction score metric. In addition, several scoring rules for probabilistic forecasts are presented to assess the interval estimates and predicted distributions, namely the logarithmic score (LogS), the quadratic score (QS), the interval score (IS), and the integrated Brier score (IBS). As pointed out by Chang and Joe (2019), scoring rules such as these are more meaningful than MAE when there is heteroscedasticity in the conditional distribution. For every observation in the test set, the conditional expectation of EAD provides the point estimate for the MAE measure.

To produce the interval scores, 95% prediction intervals bounded by the 0.025 and 0.975 quantile levels are taken as the interval estimates. The performance measures are defined as follows (Gneiting and Raftery, 2007). Firstly,

$$\text{MAE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} |y^{(i)} - \hat{y}^{(i)}|,$$

where $\hat{y}^{(i)}$ is the predicted conditional expectation of EAD. Second, the logarithmic and quadratic scores measure the quality of the predicted density (the latter incorporating an L_2 penalty term), as follows:

$$\begin{aligned} \text{LogS} &= \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \log \hat{f}_{Y|\mathbf{X}}(y^{(i)}|\mathbf{x}^{(i)}), \\ \text{QS} &= \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[2\, \hat{f}_{Y|\mathbf{X}}(y^{(i)}|\mathbf{x}^{(i)}) - \int_{-\infty}^{\infty} \hat{f}_{Y|\mathbf{X}}(y|\mathbf{x}^{(i)})^2\, dy \right], \end{aligned}$$

where $(\mathbf{x}^{(i)}, y^{(i)})$ are the actual observations and $\hat{f}_{Y|\mathbf{X}}$ is the predicted conditional PDF. Third, the interval score evaluates interval forecasts rewarding narrow prediction intervals whilst penalising observations falling outside those intervals. Specifically,

$$\text{IS} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[(\hat{u}^{(i)} - \hat{l}^{(i)}) + \frac{2}{\alpha}(\hat{l}^{(i)} - y^{(i)}) \mathbb{I}\{y^{(i)} < \hat{l}^{(i)}\} + \frac{2}{\alpha}(y^{(i)} - \hat{u}^{(i)}) \mathbb{I}\{y^{(i)} > \hat{u}^{(i)}\} \right],$$

where, for a $(1-\alpha)100\%$ prediction interval, $\hat{l}^{(i)}$ and $\hat{u}^{(i)}$ are the predicted lower and upper bounds at quantile levels $\alpha/2$ and $1-\alpha/2$, respectively. We select $\alpha = 0.05$. Lastly, the integrated Brier score provides a performance measure for the predicted cumulative distribution:

$$\text{IBS} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \int_{-\infty}^{\infty} \left[\hat{F}_{Y|\mathbf{X}}(y|\mathbf{x}^{(i)}) - \mathbb{I}\{y \geq y^{(i)}\} \right]^2 dy,$$

where $\hat{F}_{Y|\mathbf{X}}$ denotes the predicted conditional CDF.

Model	MAE ↓	LogS ↑	QS ↑	IS ↓	IBS ↓
OLS	9871	-11.34	2.19e-05	62695	7042
LQR	9322	-	-	42979	-
P-DVQR	11400	-10.48	8.81e-05	45689	8677
NP-DVQR	**8572**	**-10.04**	**9.85e-05**	**41795**	**6129**

TABLE 4.4: Performance results for point and interval estimates as well as distributions (bold face indicates best performance). The arrows indicate that lower values for MAE, IS and IBS, and higher values for LogS and QS, imply better performance.

Table 4.4 summarises the performance of all models according to these metrics. Note that, as the predictive density and cumulative distributions of the response for LQR cannot be extracted analytically, its LogS, QS and IBS were excluded. Compared with

OLS, we observe that LQR produces better point and interval estimates (see lower MAE and IS, respectively). In particular, the substantial reduction in IS confirms that the linear quantile regression model is capable of providing a much more reliable prediction interval than the linear model. However, LQR is itself outperformed by non-parametric DVQR, which yields even better point and interval estimates. In fact, NP-DVQR exhibits superior performance on all five measures, so it is the preferred method regardless of the intended model application. Again, to avoid misspecification of the dependencies, it proves important to use non-parametric DVQR, as P-DVQR shows poorer performance relative to NP-DVQR.

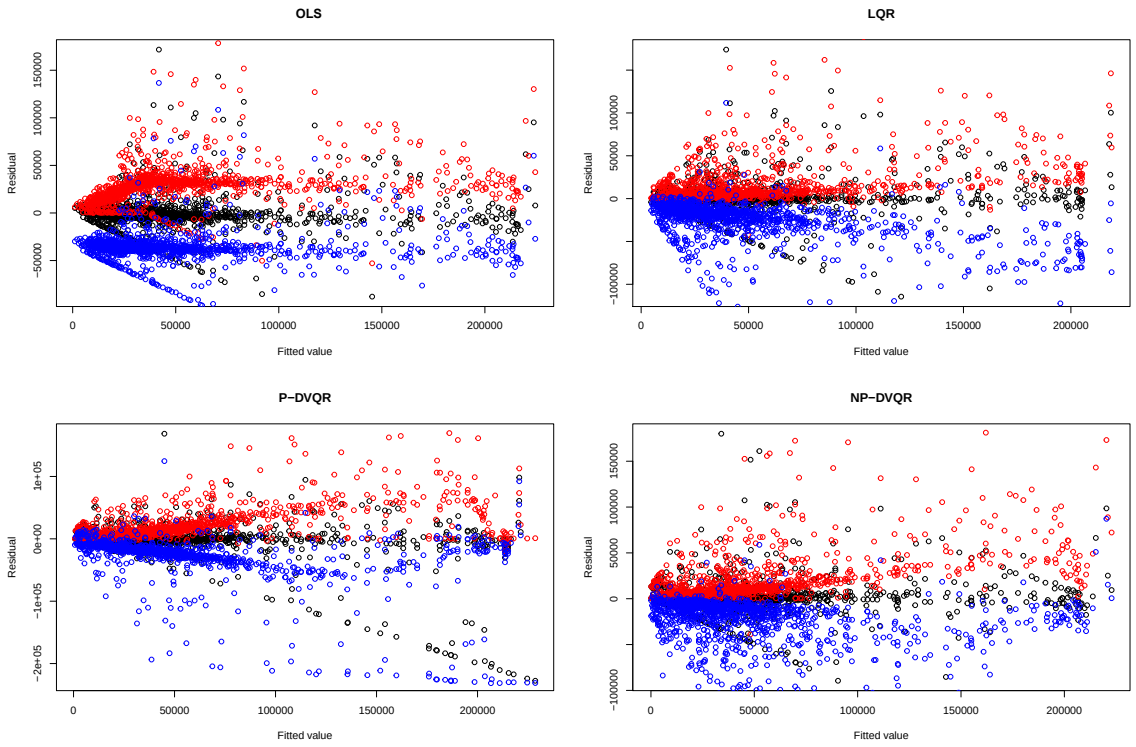


FIGURE 4.10: Residual vs. fitted plots extracted from a sample of EAD data (for clearer visualisation). Black dots denote the residuals; the red and blue dots are the lower and upper bound of the prediction intervals, respectively.

Figure 4.10 contains residual plots for all models, including the 95% model prediction intervals plotted against fitted mean values. As shown, all quantile regression models capture the conditional heteroscedasticity, producing a wider prediction interval as fitted values increase. Conversely, OLS fails to reflect this, as its prediction interval does not further widen for much of the value range. This explains why LQR and DVQR models provide more precise prediction intervals than OLS, according to the interval scores shown earlier in Table 4.4.

4.7 Conclusions and future research

Using a large dataset of credit card defaults, this paper has applied linear and D-vine copula-based quantile regression models to predict conditional quantiles of the Exposure At Default (EAD), i.e. the card balance at default time. Exploratory data analysis revealed that the marginal distributions of EAD and its covariates are non-normal, have high variance and exhibit heteroscedasticity. Hence, interval estimate models, such as quantile regression, that make no parametric distribution assumption and do not require constant variance, are generally more suitable for modelling such data than point estimate models such as OLS linear regression. Quantile regression models also have the added advantage of allowing for the variable effects to differ depending on the EAD quantile of interest. For example, our analyses have shown that the credit limit has a substantially larger impact on higher EAD quantiles (and thus tail risk) than on its mean or lower quantiles. Furthermore, we observed an improvement in the predicted conditional quantiles and the point and interval estimates for EAD when the quantile models are employed instead of the OLS model.

Among the different quantile models tried in the paper, the D-vine copula models have distinct advantages over the linear quantile model, as they address two problems that may be associated with classical quantile regression: the occurrence of quantile crossings and multicollinearity problems. Specifically, the pair-copulas fitted by the newly proposed D-vine quantile regression also produce deeper insights into the complex high-dimensional dependence structure between EAD and the covariates, as well as between the covariates themselves. We thus detected several pairwise asymmetric and tail dependencies that are overlooked by the other methods, including, for example, pronounced tail dependence between EAD and the current credit limit. Also, the method revealed non-linear and non-monotonic predictor effects at several EAD quantile levels. What's more, a predictive performance comparison on the real-life data showed that the D-vine copula quantile regression model with non-parametric copulas outperforms the other models, yielding better point and interval estimates for EAD than the linear quantile model, and more closely reflecting the actual distribution of EAD than the OLS linear model. In summary, we conclude that non-parametric D-vine copula-based quantile regression is a highly attractive approach when predictions of conditional quantiles and interval estimates for EAD are required.

A future avenue of research is to model another Basel risk parameter, namely the Loss Given Default (LGD), using vine copula-based quantile regression. Similarly to EAD data, variables in LGD datasets are often found to be correlated through asymmetric and non-linear structures, making conventional correlation analysis unsuitable. Moreover, estimating the upper tail or higher quantiles of LGD is again more relevant

for calculating unexpected losses or required capital than estimating the average value. By utilising the proposed method to model LGD, we conjecture that point and interval estimates can be similarly improved.

Chapter 5

Conclusions and future research

This three-paper thesis contributes to the consumer credit risk literature by modelling Exposure At Default (EAD) for credit card portfolios. Three novel EAD models have been developed and tested on real-life data to improve the predictive performance and gain additional interpretation insights.

In the first paper, two distinct groups of card borrowers were considered: those whose balance hits the limit in the race to default, i.e. those who "max out" their card, and those who do not. It was hypothesised that not only the level of EAD but also its risk drivers could differ substantially between these two groups. Hence, we proposed a two-component mixture model that conditions EAD on these two respective scenarios, using the GAMLSS framework. The proposed and other benchmark models were empirically validated through a series of discrimination and calibration measures. The results showed a clear performance benefit of combining the mixture component and the GAMLSS framework over the OLS models. This confirms, consistently with what Tong et al. (2016) and Leow and Crook (2016) reported for other datasets, that there are indeed predictive accuracy gains to be had in EAD modelling from including non-linear effects and targeting not only the EAD mean but also dispersion (cf. Tong et al. (2016)), as well as from distinguishing between the two max-out scenarios (cf. Leow and Crook (2016)). In terms of predictor effects, the current limit was found to be the variable with the strongest impact on the mean of (non-zero) EAD, whereas the current balance and (estimated) time to default strongly affect EAD dispersion. Furthermore, the risk drivers for the borrowers from the two groups were shown to be different. For instance, current balance was selected in the model for the max-out group, but not the other group, whereas current limit was identified as being strongly linked to dispersion only under the non-max-out scenario. Lastly, the max-out model component of the mixture model provides banks with useful insights as to the probability that a borrower will max out their credit card and the factors that contribute to this. Such model component may have further applications in allowing the bank to actively manage the credit limit of those who are most at risk.

In the second paper, a novel approach was developed to model the Probability of Default (PD), Exposure At Default (EAD), and their dependence, in a retail credit card portfolio. The rationale for doing so was that previous studies have shown that accounting for such dependence is important to avoid underestimating expected portfolio loss and, hence, the capital requirement. A joint model for PD and EAD was developed by applying the bivariate Copula Generalised Additive Models for Location, Scale and Shape (CGAMLSS) framework — the first such application of this framework in the credit risk setting. The research also explicitly addressed potential sample selection bias, by not restricting the data to just the defaulted accounts (as most EAD studies do), but extending the analysis to outstanding balance (rather than balance at default time, or equivalently, EAD) in a larger sample of both defaulted and non-defaulted accounts. This allows us to avoid potential misestimation of expected loss when the models are applied to an entire portfolio of accounts, not only the subsample of those that will default. To empirically validate the effectiveness of introducing the dependence, the newly proposed copula model was benchmarked against two standalone models, for PD and balance, which were separately constructed and not considering such dependence. According to our dataset, the analysis showed that accounts with higher default likelihood tended to end up having a higher card balance. More specifically, a strong and positive dependence between PD and balance was revealed, even after accounting for observable covariates, either in the middle (Frank copula) or upper tail (180°Clayton copula) area of the marginal distributions. The distribution of future balance also showed a heavier tail for accounts that are more likely to default. In addition, a series of significant risk factors was identified: credit rating provided the largest impact on PD, future balance was most strongly influenced by current balance, and their dependence was most affected by credit utilisation. Moreover, the proposed CGAMLSS model produced more accurate and conservative expected loss estimates which, in agreement with previous literature findings, are exhibiting a heavy tail that is the result of the correlation between PD and credit card balance. Lastly, by, instead, ignoring such dependence or by allowing sample selection bias, loss could be severely underestimated, on our dataset, by a percentage error of 10.82% and 5.17%, respectively. This potentially leads to substantial capital shortfalls. We found the proposed copula models to be better as they overestimated the loss with smaller percentage errors of 1.16% (180°Clayton copula) and 2.14% (Frank copula).

The third paper is the first to estimate conditional mean and quantiles and interval estimates for EAD using a state-of-the-art quantile regression method — vine copula-based quantile regression. We argued that this approach has several benefits. First, similarly to other quantile regression methods, it provides further insights on the right tail area of EAD distribution, i.e. higher quantiles of EAD, which is useful for risk management and capital calculation. Second, the vine copula approach in particular can be applied to model multi-dimensional dependencies among all variables in an

EAD dataset, through a suitable series of (either parametric or non-parametric) pair-copulas. In so doing, it avoids the multicollinearity problems faced by classical quantile regression. Another benefit of the vine copula approach is that, unlike with classical quantile regression, it guarantees that the regression lines of different quantile levels do not cross each other, which facilitates their interpretation. Using a real-life dataset of credit card accounts, our analysis further showed that the proposed model with non-parametric copulas produced better predictive point and interval estimates for EAD than conventional linear quantile model, and that it more closely reflected the actual distribution of EAD compared to other models. In addition, the approach was able to identify several pairwise asymmetric and tail dependencies between EAD and the input variables, as well as between the input variables themselves, which would otherwise be overlooked; for example, pronounced positive upper tail dependence between EAD and the current credit limit was detected. This implies that additional capital may be required for accounts with a higher limit, to cope with their higher tail risk. Lastly, in our dataset, the estimated parametric copulas could not correctly model the non-monotonic relationships between the variables, leading to worse performance compared to when the non-parametric copulas were employed.

In summary, the GAMLSS framework has proved an attractive approach to model the challenging distribution of EAD (or, more generally, credit card balance) because it offers a wide range of options for parametric distributions which are not restricted to the exponential family. Specifically, we recommend a zero-adjusted gamma distribution when the observed EAD data contain several zero values and exhibit positive skewness. Using this framework, non-monotonic relationships between the response's parameters and predictors can also be effectively modelled by non-parametric splines. Second, the copula approach was found to perform effectively in modelling the dependence between PD and balance and revealed their asymmetric right tail dependence structure. Also, the estimated loss showed a heavier tail when such dependence was considered. Based on these findings, copula regression could provide an alternative method, for practitioners, to calculate required capital, in a manner that is more conservative and, possibly, more accurately reflects the tail risk of loss. Instead, in the scenario where the dependence between PD and EAD is not modelled, the third paper allows one to produce EAD estimates that also incorporate a larger margin of conservatism, by employing higher EAD quantiles rather than the mean level. This conforms with the Basel regulation which requires more conservative values for EAD when such dependence is not considered in the model. Lastly, when we consider the predictor effects found across the three papers, the GAMLSS models in the first paper suggested several non-monotonic relationships which could possibly occur because of sample selection bias. However, in the second paper, where such bias was addressed, these non-monotonic effects still persisted. In the third paper, though, we found that more of them are now modelled as being monotonic. One possible explanation could be that the non-monotonic effects observed by the former methods

might be (in part) linked to multicollinearity which was alleviated by the use of vine copulas in the third paper.

Potential avenues of future research, which follow from the work in the thesis, are as follows. Whilst the first paper showed the benefits of adding a mixture component to the EAD model, the second paper utilised the copula method to better capture the dependence between PD and credit card balance. Therefore, a further extension to the work would be to combine both methods and jointly model PD and balance, conditioned on whether a max-out event occurs. We conjecture that the accounts whose balance hits the limit have both a higher chance to default and also a higher future balance level. Moreover, in so doing, we may gain additional insights on how the predictor effects on PD, balance, and their dependence structure, may vary in each group. Ultimately, we propose to test whether this extension could further improve the estimation of expected loss.

Secondly, although the second paper considered both PD and EAD, and their dependence, it ignored LGD and its relationship with the other two IRB risk parameters. Therefore, further research could seek to model all three parameters, and the dependencies between them. One solution may be to extend the application of the CGAMLSS framework, used in the second paper, to a trivariate analysis. An alternative approach may be to model PD, LGD and EAD using the D-vine copulas. Both methods, however, share the common challenge that LGD is not observed for the non-defaults (unlike EAD where outstanding balance could be used instead).

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44:182–198.

Adrian, T. and Brunnermeier, M. K. (2016). Covar. *American Economic Review*, 106(7):1705–41.

Agarwal, S., Ambrose, B., and Liu, C. (2006). Credit lines and credit utilization. *Journal of Money, Credit, and Banking*, 38:1–22.

Allen, L. and Saunders, A. (2003). A survey of cyclical effects in credit risk measurement model. BIS Working Papers 126, Bank for International Settlements.

Altman, E. I., Brady, B., Resti, A., and Sironi, A. (2005). The link between default and recovery rates: Theory, empirical evidence, and implications. *The Journal of Business*, 78:2203–2228.

Apostolik, R., Donohue, C., and Went, P. (2009). *Foundations of banking risk : an overview of banking, banking risks, and risk-based banking regulation*. John Wiley and Sons.

Araten, M. and Jacobs, M. (2001). Loan equivalents for revolving credits and advised lines. *The RMA Journal*, 83:34–39.

Bade, B., Rösch, D., and Scheule, H. (2010). Default and recovery risk dependencies in a simple credit risk model. *European Financial Management*, 17:120–144.

Bager, A. (2018). Ridge parameter in quantile regression models. an application in biostatistics. *International Journal of Statistics and Applications*, 8:72–78.

Balakrishnan, N. and Lai, C. D. (2009). *Continuous bivariate distributions*. New York Springer Cop.

Banasik, J., Crook, J. N., and Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, 50:1185–1190.

Barakova, I. and Parthasarathy, H. (2013). Modeling corporate exposure at default. Available at SSRN: https://ssrn.com/abstract=2235218.

Barco, M. (2007). Going downturn. *Risk*, 20:39–44.

BCBS (1999). *Credit Risk Modelling: Current Practices and Applications*. Basel Committee on Banking Supervision.

BCBS (2017). *Basel III: Finalising post-crisis reforms*. Bank for International Settlements.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268. Cited By :421.

Bedford, T. and Cooke, R. M. (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30:1031–1068.

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60:1699–1707.

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

Bernard, C. and Czado, C. (2015). Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, 138:104–126.

Bouyé, E. and Salmon, M. (2009). Dynamic copula quantile regressions and tail area dynamic dependence in forex markets. *The European Journal of Finance*, 15:721–750.

Brown, I. (2011). *Regression Model Development for Credit Card Exposure At Default (EAD) using SAS/STAT® and SAS® Enterprise Miner™ 5.3*. SAS Global Forum, Las Vegas, NV.

Calabrese, R., Osmetti, S. A., and Zanin, L. (2019). A joint scoring model for peer-to-peer and traditional lending: a bivariate model with copula dependence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182:1163–1188.

Caselli, S., Gatti, S., and Querci, F. (2008). The sensitivity of the Loss Given Default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Services Research*, 34:1–34.

Chang, B. and Joe, H. (2019). Prediction based on conditional distributions of vine copulas. *Computational Statistics and Data Analysis*, 139:45–63.

Chava, S., Stefanescu, C., and Turnbull, S. (2011). Modeling the loss distribution. *Management Science*, 57:1267–1287.

Conn, A. R., Gould, N. I. M., and Toint, P. L. (2000). *Trust-region methods*. Siam.

Dette, H., Van Hecke, R., and Volgushev, S. (2014). Some comments on copula-based regression. *Journal of the American Statistical Association*, 109:1319–1324.

Duong, T. (2016). Non-parametric smoothed estimation of multivariate cumulative distribution and survival functions, and receiver operating characteristic curves. *Journal of the Korean Statistical Society*, 45:33–50.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121.

Embrechts, P., Lindskog, F., and Mcneil, A. (2003). Modelling dependence with copulas and applications to risk management. In Rachev, S. T., editor, *Handbook of Heavy Tailed Distributions in Finance*, volume 1, chapter 8, pages 329 – 384. North-Holland.

Frye, J. (2000). Collateral damage. *Risk*, 13:91–94.

Geidosch, M. and Fischer, M. (2016). Application of vine copulas to credit portfolio risk modeling. *Journal of Risk and Financial Management*, 9:1–15.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12:347–368.

Gibilaro, L. and Mattarocci, G. (2018). Multiple banking relationships and exposure at default. *Journal of Financial Regulation and Compliance*, 26(1):2–19.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gu, C. (1992). Cross-validating non-gaussian data. *Journal of Computational and Graphical Statistics*, 1:169.

Gürtler, M., Hibbeln, M. T., and Usselmann, P. (2018). Exposure at default modeling – a theoretical and empirical assessment of estimation approaches and parameter choice. *Journal of Banking and Finance*, 91:176–188.

Hahn, R., Reitz, S., Engelmann, B., and Rauhmeier, R. (2011). Possibilities of estimating exposures. In *The Basel II Risk Parameters: Estimation, Validation, Stress Testing - with Applications to Loan Risk Management*, page 185–200. Springer, 2 edition.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1:297–318.

Haupt, H., Kagerer, K., and Schnurbus, J. (2011). Cross-validating fit and predictive accuracy of nonlinear quantile regressions. *Journal of Applied Statistics*, 38:2939–2954.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Hillebrand, M. (2006). Modeling and estimating dependent loss given default. *Risk*, 19:120–125.

Hon, P. S. and Bellotti, T. (2016). Models and forecasts of credit card balance. *European Journal of Operational Research*, 249(2):498–505.

Hosmer, D., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. Wiley, 3 edition.

Jacobs, M. (2010). An empirical study of exposure at default. *Journal of Advanced Studies in Finance*, 1(1):31–59.

Jacobs, M. and Bag, P. (2010). What do we know about exposure at default on contingent credit lines? - a survey of the literature, empirical analysis and models. *Journal of Advanced Studies in Finance*, II:26–46.

Jacobs, M. and Karagozoglu, A. K. (2011). Modeling ultimate Loss Given Default on corporate debt. *The Journal of Fixed Income*, 21:6–20.

Jiménez, G., Lopez, J. A., and Saurina, J. (2009). Empirical analysis of corporate credit lines. *Review of Financial Studies*, 22:5069–5098.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In Rüschendorf, L., Schweizer, B., and Taylor, M. D., editors, *Distributions with fixed marginals and related topics*, volume 28 of *Lecture Notes–Monograph Series*, pages 120–141. Institute of Mathematical Statistics, Hayward, CA.

Kaposty, F., Löderbusch, M., and Maciag, J. (2017). Stochastic loss given default and exposure at default in a structural model of portfolio credit risk. *The Journal of Credit Risk*, 13(1):93–123.

Kauermann, G. and Schellhase, C. (2013). Flexible pair-copula estimation in d-vines using bivariate penalized splines. *Statistics and Computing*, 24:1081–1100.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in germany. *The Annals of Applied Statistics*, 9:1024–1052.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 46:33–50.

Koenker, R. and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94:1296–1310.

Komunjer, I. (2013). Quantile Prediction. In *Handbook of Economic Forecasting*, volume 2, pages 767–785. Elsevier.

Kraus, D. and Czado, C. (2017). D-vine copula based quantile regression. *Computational Statistics and Data Analysis*, 110:1–18.

Krüger, S., Oehme, T., Rösch, D., and Scheule, H. (2018). A copula sample selection model for predicting multi-year LGDs and lifetime expected losses. *Journal of Empirical Finance*, 47:246–262.

Krüger, S. and Rösch, D. (2017). Downturn LGD modeling using quantile regression. *Journal of Banking and Finance*, 79:42–56.

Krämer, N., Brechmann, E. C., Silvestrini, D., and Czado, C. (2013). Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53:829–839.

Kupiec, P. H. (2008). A generalized single common factor model of portfolio credit risk. *The Journal of Derivatives*, 15:25–40.

Lam, J. (2014). *Enterprise risk management: From incentives to controls*. John Wiley and Sons, Ltd, 2 edition.

Leow, M. and Crook, J. (2016). A new mixture model for the estimation of credit card exposure at default. *European Journal of Operational Research*, 249(2):487–497.

Leow, M. and Mues, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, 28(1):183–195.

Luo, S. and Murphy, A. (2020). Understanding the Exposure at Default Risk of Commercial Real Estate Construction and Land Development Loans. Working Papers 2007, Federal Reserve Bank of Dallas. Available at https://ideas.repec.org/p/fip/feddwp/87677.html.

Malik, M. and Thomas, L. C. (2007). Modeling credit risk of portfolio of consumer loans. *SSRN Electronic Journal*.

Malik, M. and Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society*, 61:411–420.

Marra, G. and Radice, R. (2017a). Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, 112:99–113.

Marra, G. and Radice, R. (2017b). *GJRM: generalized joint regression modelling*.

Martey, E. N. and Attoh-Okine, N. (2019). Analysis of train derailment severity using vine copula quantile regression modeling. *Transportation Research Part C: Emerging Technologies*, 105:485–503.

Mashal, R. and Zeevi, A. (2002). Beyond correlation: Extreme co-movements between financial assets. *SSRN Electronic Journal*.

Mccullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, 2 edition.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29:449–470.

Mester, L. J., Nakamura, L. I., and Renault, M. (2006). Transactions accounts and loan monitoring. *Review of Financial Studies*, 20:529–556.

Miu, P. and Ozdemir, B. (2006). Basel requirements of downturn loss given default: modeling and estimating probability of default and loss given default correlations. *The Journal of Credit Risk*, 2:43–68.

Moral, G. (2006). EAD estimates for facilities with explicit limits. In Engelmann, B. and Rauhmeier, R., editors, *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing*, pages 197–242. Springer Berlin Heidelberg, Berlin, Heidelberg.

Moreira, F. F. (2010). Copula-based formulas to estimate unexpected credit losses (the future of Basel Accords?). *Financial Markets, Institutions and Instruments*, 19:381–404.

Nagler, T. and Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89.

Nagler, T. and Kraus, D. (2019). *vinereg: D-Vine Quantile Regression*.

Nagler, T., Schellhase, C., and Czado, C. (2017). Nonparametric estimation of simplified vine copula models: comparison of methods. *Dependence Modeling*, 5:99–120.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Publishing Company, Incorporated, 2 edition.

Niemierko, R., Töppel, J., and Tränkler, T. (2019). A d-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data. *Applied Energy*, 233-234:691–708.

Nikoloulopoulos, A. K., Joe, H., and Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics and Data Analysis*, 56:3659–3673.

Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer.

Norden, L. and Weber, M. (2010). Credit line usage, checking account activity, and default risk of bank borrowers. *Review of Financial Studies*, 23:3665–3699.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076.

Pluto, K. and Tasche, D. (2011). Estimating probabilities of default for low default portfolios. In Engelmann, B. and Rauhmeier, R., editors, *The Basel II Risk Parameters:*

Estimation, Validation, Stress Testing - with Applications to Loan Risk Management, pages 75–101. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pykhtin, M. (2003). Unexpected recovery risk. *Risk*, 16:74–78.

Qi, M. (2009). Exposure at default of unsecured credit cards. Economics working paper 2009-2, Office of the Comptroller of the Currency. Available at https://www.occ.gov/publications-and-resources/publications/economics/working-papers/pub-econ-working-paper-2009-2.pdf.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radice, R., Marra, G., and Wojtyś, M. (2015). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26:981–995.

Rigby, B. and Stasinopoulos, M. (2010). *A flexible regression approach using GAMLSS in R*.

Rösch, D. and Scheule, H. (2008). Downturn LGD for Hong Kong mortgage loan portfolios. *The Journal of Risk Model Validation*, 2:3–11.

Rösch, D. and Scheule, H. (2014). Forecasting probabilities of default and loss rates given default in the presence of selection. *Journal of the Operational Research Society*, 65:393–407.

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41.

Savu, C. and Trede, M. (2009). Hierarchies of archimedean copulas. *Quantitative Finance*, 10:295–304.

Schallhorn, N., Kraus, D., Nagler, T., and Czado, C. (2017). D-vine quantile regression with discrete variables.

Scheffer, M. and Weiß, G. N. F. (2016). Smooth nonparametric bernstein vine copulas. *Quantitative Finance*, 17:139–156.

Schirmacher, D. and Schirmacher, E. (2008). Multivariate dependence modeling using pair-copulas. Technical report, Society of Acturaries: 2008 Enterprise Risk Management Symposium, April 14-16, Chicago.

Sigrist, F. and Stahel, W. A. (2011). Using the censored gamma distribution for modeling fractional response variables with an application to Loss Given Default. *ASTIN Bulletin*, 41(2):673–710.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Somers, M. and Whittaker, J. (2007). Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, 183:1477–1487.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: Using GAMLSS in R*. Chapman and Hall.

Stasinopoulos, M. D., Rigby, R. A., Voudouris, V., Akantziliotou, C., Enea, M., and Kiose, D. (2019). *Generalised Additive Models for Location Scale and Shape*.

Stöber, J., Joe, H., and Czado, C. (2013). Simplified pair copula constructions—limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118.

Taplin, R., To, H. M., and Hee, J. (2007). Modeling exposure at default, credit conversion factors and the Basel II Accord. *The Journal of Credit Risk*, 3(2):75–84.

Thackham, M. and Ma, J. (2018). Exposure at default without conversion factors – evidence from global credit data for large corporate revolving facilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1267–1286.

Thomas, L. C., Crook, J. N., and Edelman, D. B. (2017). *Credit scoring and its applications*. Society for Industrial and Applied Mathematics.

Tong, E. N., Mues, C., Brown, I., and Thomas, L. C. (2016). Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research*, 252(3):910–920.

Tong, E. N., Mues, C., and Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, 29:548–562.

Tong, E. N., Mues, C., and Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, 218:132–139.

Trivedi, P. K. and Zimmer, D. M. (2006). Copula modeling: An introduction for practitioners. *Foundations and Trends® in Econometrics*, 1:1–111.

Valvonis, V. (2008). Estimating EAD for retail exposures for Basel II purposes. *The Journal of Credit Risk*, 4:79–109.

Van Gestel, T., Baesens, B., Van Dijcke, P., Garcia, J., Suykens, J. A., and Vanthienen, J. (2006). A process model to develop an internal rating system: Sovereign credit ratings. *Decision Support Systems*, 42:1131–1151.

Vasicek, O. (2002). Loan portfolio value. *Risk.net*, (2):160–162.

Vatter, T. and Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167.

Witzany, J. (2011). Exposure at default modeling with default intensities. *European Financial and Accounting Journal*, 6(4):20–48.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99:673–686.

Yang, B. H. and Tkachenko, M. (2012). Modeling exposure at default and loss given default: empirical approaches and technical implementation. *The Journal of Credit Risk*, 8(2):81–102.

Yee, T. W. (2016). *Vector Generalized Linear and Additive Models : with an implementation in r*. Springer.