Journal Pre-proof

Modelling Credit Card Exposure At Default Using Vine Copula Quantile Regression

Suttisak Wattanawongwan, Christophe Mues, Ramin Okhrati, Taufiq Choudhry, Mee Chi So

PII: S0377-2217(23)00380-6
DOI: https://doi.org/10.1016/j.ejor.2023.05.016
Reference: EOR 18483

To appear in: *European Journal of Operational Research*

Received date: 7 March 2022
Accepted date: 8 May 2023

Please cite this article as: Suttisak Wattanawongwan, Christophe Mues, Ramin Okhrati, Taufiq Choudhry, Mee Chi So, Modelling Credit Card Exposure At Default Using Vine Copula Quantile Regression, *European Journal of Operational Research* (2023), doi: https://doi.org/10.1016/j.ejor.2023.05.016

Highlights

- We use vine copula-based quantile regression to produce EAD quantile estimates.
- Our vine copulas effectively model complex dependencies among the predictors for EAD.
- We provide insights into how the predictor effects vary per EAD quantile.
- The proposed model outperforms a linear quantile model on real-life credit card data.

Modelling Credit Card Exposure At Default Using Vine Copula Quantile Regression

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

[a] *School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK*
[b] *Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, 10400, Thailand*
[c] *Southampton Business School, University of Southampton, Highfield, Southampton, SO17 1BJ, UK*
[d] *Institute of Finance and Technology, University College London, London, WC1E 6BT, 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 data and its non-normal empirical distribution, 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 vine copula-based quantile regression, an interval estimation approach, to model the entire distribution of EAD and predict its conditional mean and quantiles. This methodology addresses several drawbacks of classical quantile regression, including quantile crossing and multicollinearity, and it 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 estimates for EAD, and more accurately reflects its actual distribution, compared to a selection of other models.

Keywords: Risk analysis, Credit cards, Exposure At Default, Quantile regression, Vine copulas

1. Introduction

Under the Advanced Internal Ratings-Based (A-IRB) approach, the Basel II and III Accords allow authorised banks to calculate 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

*Corresponding author
Email address: suttisak.wat@mahidol.ac.th (Suttisak Wattanawongwan)

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 upper 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 methods 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), Leow and Crook (2016), and Wattanawongwan et al. (2023) 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 large Asian retail lender. For most A-IRB banks, credit cards account for the largest number of defaults, which are often scarce in practice among other 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 each quantile model 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 2, from which the main contributions of the paper are then identified. Section 3 explains the data and variables used, and Section 4 provides a brief description of vine copulas. Section 5 illustrates how the statistical models are constructed. The results are analysed in Section 6. Section 7 concludes.

2. Literature review

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

2.1. EAD modelling

For revolving credit 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 (see, for example, Bellotti and Crook (2012)), and random effects panel models (Bollen and Brand, 2010). 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. Wattanawongwan et al. (2023) later added a similar mixture component to their GAMLSS models, finding that it further improved the predictive performance. None of these methods explicitly studied interval estimates, although Tong et al. (2016) and Wattanawongwan et al. (2023) did model a dispersion parameter.

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; 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 having several pitfalls. Kraus and Czado (2017) 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, which, vice versa, is constrained by the former. 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); Geidosch and Fischer (2016)).

2.3. Copulas

Copulas are a more appropriate method to model complex dependence patterns (for standard references on copula theory, we refer the reader to the books by Joe (1997) and Nelsen (2006)). Over the past decades, copulas have become increasingly popular in finance and insurance settings (Nelsen, 2006; Krüger et al., 2018; Calabrese et al., 2019). 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 (governed by the same parameter). On the other hand, the Gaussian copula has no tail dependence.

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).

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 subse-

quently developed: the Canonical vine (C-vine) and the Drawable vine (D-vine) (Aas et al., 2009). They have been applied actively in financial and insurance risk management; see, for example, Nikoloulopoulos et al. (2012). For a more comprehensive treatment of vine copulas, we refer the reader to Czado (2019).

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 an 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 forward variable selection. 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. More recently, Tepegjozova et al. (2022) introduced a two-step ahead forward selection algorithm and extended the approach to be applicable to both C-vine and D-vine copulas. The analyses in our paper are restricted, however, to D-vine copulas and do not consider other vine types, such as C-vines or the more general class of R-vines (Zhu et al., 2021). Although R-vines may offer increased flexibility in modeling large-volume, high-dimensional data (Dissmann et al., 2013), they produce a huge number of possible vine structures, and, hence, have seen fewer applications in practice (Yu et al., 2020). C-vines are used to fit a multi-variable model with a key variable that governs the dependencies and interactions in the dataset (Aas et al., 2009; Yu et al., 2020). D-vine models, having been more widely used than C-vine models (Martey and Attoh-Okine, 2019), may be preferred to C-vines when one does not want to assume a key variable that controls the dependencies and little prior knowledge exists on the dependence structures between variables (Yu et al., 2020). For these reasons, and since our objective is to find the unknown (nonlinear/non-monotonic) relationships between variables, the paper excludes C-vines and R-vines, and focuses instead on D-vines.

Although vine copulas have recently been used in other settings such as inventory financing

(Zhi et al., 2020), to the best of the authors' knowledge, this paper is the first to propose the vine copula-based quantile regression framework in any consumer credit risk setting.

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 account-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.

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 Asian 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 1st 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. Note that data from the same lender has also been used by Wattanawongwan et al. (2023) to empirically test a series of mixture models which, unlike the current paper, do not produce interval estimates or consider the dependencies between the variables.

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 1: List of available explanatory variables.

Table 1 lists the explanatory variables; all of these are continuous and 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 more than 60,000. 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, the interquartile range, $\text{IQR} = Q_3 - Q_1$, with Q_1 and Q_3 denoting the lower and upper quartile, respectively (Dekking et al., 2005).

For an exploratory analysis of the distribution of each variable and their bivariate relationships (revealing some asymmetric and tail dependencies that the vine copula quantile regression approach should be capable of handling), we refer the reader to the Supplementary Materials, Online Appendix A.

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 Sklar'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:

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

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 density $c(u_1, \ldots, u_p)$ can be decomposed into a product of $\frac{p(p-1)}{2}$ (conditional) bivariate copula densities, or the so-called pair-copula densities (Bedford and Cooke, 2001). Following Aas et al. (2009), a D-vine Pair-Copula Construction (PCC) with order $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_p$ of the joint density f can be written as:

$$\begin{aligned} f(x_1 \ldots, x_p) = \prod_{k=1}^{p} f_k(x_k) \prod_{i=1}^{p-1} \prod_{j=i+1}^{p} c_{ij|i+1,\ldots,j-1}(F_{i|i+1,\ldots,j-1}(x_i|x_{i+1}, \ldots, x_{j-1}), \\ F_{j|i+1,\ldots,j-1}(x_j|x_{i+1}, \ldots, x_{j-1})|x_{i+1}, \ldots, x_{j-1}), \end{aligned} \tag{2}$$

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 $F_{j|D}(x_j|X_D = x_D)$.

To enable fast, robust, and tractable inference, especially in higher dimensions (Haff et al., 2010; Stöber et al., 2013), we employ a simplifying assumption for the pair-copulas, i.e. that $c_{ij|D}$ does not depend on the conditioning vector X_D, i.e. $c_{ij|D}(\cdot, \cdot|x_D) = c_{ij|D}(\cdot, \cdot)$. Haff et al. (2010) argued that, even when this simplifying assumption is not completely satisfied, using such an assumption provides a good approximation. Having further analysed simulated and real data applications, Killiches et al. (2016) suggested that it might actually be beneficial for practical applications to use simplified vine copulas (i.e. making the simplifying assumption) since these could not only capture the main dependence features of the data, similarly to how non-simplified vine copulas can, but they also offer a smoother fit. Especially for models with higher dimensionality, non-simplified vine copulas can cause numerical intractability and overfitting issues. Note as well that the same simplifying assumption has been made by several previous papers applying vine copulas such as Tepegjozova et al. (2022); Martey and Attoh-Okine (2019); Kraus and Czado (2017); Schallhorn et al. (2017).

If all marginal distributions are uniformly distributed, the PCC is called a D-vine copula. We exemplify a joint multivariate density for a four-dimensional D-vine copula with order $X_1 \rightarrow X_2 \rightarrow$

$X_3 \to X_4$:

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

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).

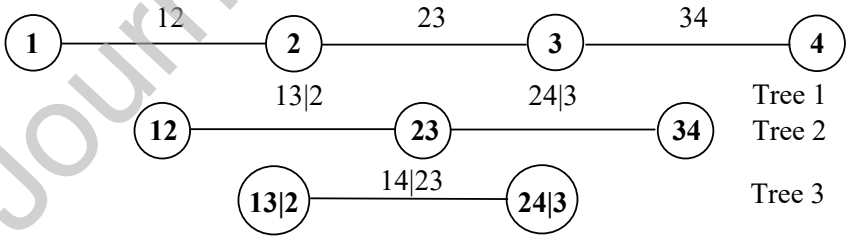


Figure 1: A four-dimensional D-vine copula with order $X_1 \to X_2 \to X_3 \to X_4$; each edge represents a pair-copula.

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 1 displays a four-dimensional D-vine structure from Equation (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 density into a product of bivariate (conditional) copula densities because the initial tree, T_1, can determine the entire structure (Kraus and Czado, 2017).

The conditional distributions $F_{i|D}(x_i|x_D)$ in Equation (2) can be estimated recursively based on pair-copulas from the respective lower trees (Joe, 1997), as follows:

$$F_{i|D}(x_i|x_D) = h_{il|D_{-l}}(F_{i|D_{-l}}(x_i|x_{D_{-l}}), F_{l|D_{-l}}(x_l|x_{D_{-l}})), \tag{4}$$

where $l \in D$ and $D_{-l} \coloneqq D \backslash \{l\}$, 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 1) can be evaluated via the h-functions related to $C_{13|2}$, C_{12}, and C_{23} from the first two trees:

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

Hence, Equation (4) allows us to estimate the joint multivariate density, $f(x_1 \ldots, x_p)$, in Equation (2), from the marginal univariate distributions, $F_1, \ldots, F_p$, and conditional pair-copulas, C_{ij}.

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.

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:

$$q_\alpha(x_1, \ldots, x_p) \coloneqq F^{-1}_{Y|X_1,\ldots,X_p}(\alpha|x_1, \ldots, x_p), \tag{5}$$

where $F_{Y|X_1,\ldots,X_p}$ is the multivariate joint distribution of $Y, X_1, \ldots, X_p$ established from a D-vine copula. By using Sklar'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 (5) can be expressed as follows (Kraus and Czado, 2017):

$$q_\alpha(x_1,\ldots,x_p) = F_Y^{-1}(C^{-1}_{V|U_1,\ldots,U_p}(\alpha|u_1,\ldots,u_p)) = F_Y^{-1}(C^{-1}_{V|U_1,\ldots,U_p}(\alpha|F_1(x_1),\ldots,F_p(x_p))). \quad (6)$$

Hence, the conditional quantile can be derived by estimating the univariate distributions F_Y and F_j and the $(p+1)$-dimensional copula $C_{V,U_1,\ldots,U_p}$. 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. More specifically, they applied a continuous convolution approach which transforms discrete variables into continuous variables by adding a small amount of noise. Nagler et al. (2017) demonstrated that this method results in a valid estimator of discrete-continuous quantile functions.

The conditional quantile, shown in Equation (6), can be extracted analytically by applying the recursion in Equation (4) and expressing $C_{V,U_1,\ldots,U_p}$ 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

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

Since $C^{-1}_{V|U_1,\ldots,U_p}(\alpha|u_1,\ldots,u_p)$ 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 estimation of the variable distributions F_Y and F_j and the copula $C_{V,U_1,\ldots,U_p}$ 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 $(x^{(1)},\ldots,x^{(n)}) \in \mathbb{R}^n$, 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 $K(x) \coloneqq \int_{-\infty}^{x} k(t)\,dt$ with $k(\cdot)$ being a symmetric probability density function and $h > 0$ a bandwidth parameter (Parzen, 1962). Following Kraus and Czado (2017), we determine the value of bandwidth h by using the plugin bandwidth from Equation (4) in Duong (2016), which minimises the asymptotic mean integrated squared error. 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: $\hat{v}^{(i)} \coloneqq \hat{F}_Y(y^{(i)})$ and $\hat{u}_j^{(i)} \coloneqq \hat{F}_j(x_j^{(i)}), j = 1,\ldots,p, i = 1,\ldots,n$.

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. cll= $\sum_{i=1}^{n} \log c_{V|U_1,\ldots,U_p}(v^{(i)}|u_1^{(i)},\ldots,u_p^{(i)})$, 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, during this procedure, a bivariate copula selection is performed based on the Akaike Information Criterion (AIC). More specifically, when a new covariate $U_k, k = 2, 3, \ldots, p$, is being added to the current D-vine copula with order $V \to U_1 \to \ldots \to U_{k-1}$, the AIC-optimal pair-copulas and their parameters (Genest and Favre) are selected from different choices of bivariate copulas. For parametric copula selection, the AIC-corrected conditional log-likelihood (cll^{AIC}) is computed as: $\text{cll}^{AIC} = -2\text{cll} + 2|\hat{\theta}|$, thus penalising the number of copula parameters, $|\hat{\theta}|$. This number of parameters is taken as the degrees of freedom for P-DVQR, and provides valuable information on its complexity. For non-parametric copula selection, we follow the computation of Nagler et al. (2017), who defined $\text{cll}^{AIC} = -2\text{cll} + 2\text{df}_e + \frac{2\text{df}_e(\text{df}_e+1)}{n-\text{df}_e-1}$, where df_e is the effective degrees of freedom (EDF) (please refer to Kauermann and Schellhase (2013), and Section 5.3.2 in Loader (2006), for explicit formulas for the EDF). Therefore, both AIC criteria include a penalty term for copula model complexity. Note that the degrees of freedom do not take the complexity of the kernel estimators for marginal CDFs into account, as they do not matter for the copula selection.

This process thus determines the pair-copulas between the response and the new covariate, $\hat{C}_{V,U_k|U_1,\ldots,U_{k-1}}$, as well as those among the existing covariates and the new covariate, $\hat{C}_{U_1,U_k|U_2,\ldots,U_{k-1}}, \hat{C}_{U_2,U_k|U_3,\ldots,U_{k-1}}, \ldots, \hat{C}_{U_{k-1},U_k}$. 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). The non-parametric copulas were estimated by a kernel estimator using local polynomial likelihoods of degree q (Geenens et al., 2017) and the corresponding $q \times q$ bandwidth matrix (see page 11 in Nagler et al. (2017)), which controls the degree of smoothing. 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 (6).

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. For extensive details, see Kraus and Czado (2017).

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. $\hat{q}_\alpha(x_1^{(i)}, \ldots, x_p^{(i)}) \coloneqq \hat{\beta}_0(\alpha) + \sum_{j=1}^{p} \hat{\beta}_j(\alpha) x_j^{(i)}$, 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 $\rho_\alpha(u) = u(\alpha - \mathbb{I}(u < 0))$ is an asymmetric loss or check function 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).

5.3. Linear regression

We introduce the OLS linear regression model (referred to as OLS) as a benchmark model with the formulation $\hat{Y}|X_1^{(i)}, \ldots, X_p^{(i)} \coloneqq \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j^{(i)} + \epsilon^{(i)}$. Here, the errors $\epsilon^{(i)}$ are assumed to be independent of each other and normally distributed with zero mean and constant variance σ^2. 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 $\hat{q}_\alpha(x_1^{(i)}, \ldots, x_p^{(i)}) \coloneqq \Phi^{-1}(\alpha|\mu^{(i)}, \sigma^2)$, where Φ^{-1} is the inverse normal CDF.

6. Analyses and results

In this section, we present the following model results: the vine copula structures, the predictor effects on EAD, EAD quantile distributions, and predictive performance. For an additional analysis of the OLS and LQR model parameter estimates, we refer the reader to Online Appendix B.

6.1. Vine copula dependence structure

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

Figure 2 exhibits the estimated D-vine copula with parametric copulas (P-DVQR), where each row represents a tree and its respective edges, with the first tree located at the bottom. The

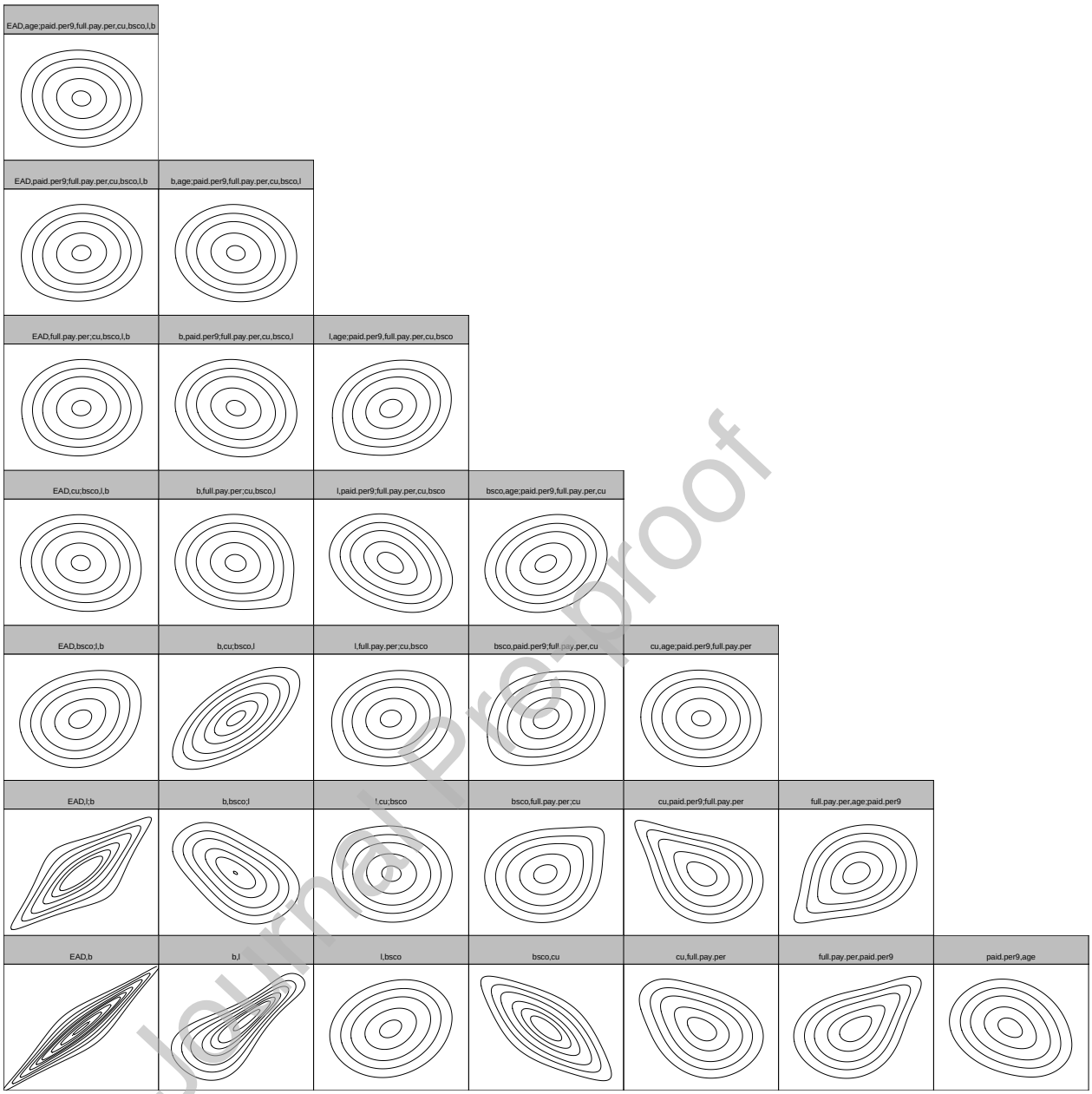


Figure 2: Estimated D-vine copula with parametric copulas and contour plots displaying the joint PDF of variable pairs with the first component on x-axis and the second on y-axis. The scale of both axes is (-3,3). The order of the D-vine copula is EAD $\rightarrow$ b $\rightarrow$ l $\rightarrow$ bsco $\rightarrow$ cu $\rightarrow$ full.pay.per $\rightarrow$ paid.per9 $\rightarrow$ age.

chosen AIC-optimal pair-copulas result in the presented contour plots, reflecting the joint PDF of the variable pair. For further detail on their maximum likelihood estimates and Kendall's tau, we refer the reader to Online Appendix C, Table C.1. The bottom row of Figure 2 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 underlying dependence between EAD and balance (see the plot in the bottom-left corner of Figure 2) is strongly positive and symmetric, exhibiting both upper and lower tail dependence (as implied by the choice of the Student-t copula, see Table C.1). 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 (second to bottom row of Figure 2) is also captured by the t copula but here the correlation is weaker. Rating score shows some correlation with EAD as well, with a mild positive upper tail dependence captured by the Joe-Frank (BB8) copula. The dependencies between EAD and the other covariates are relatively weak. Several strongly related variable pairs are also found among the explanatory variables themselves. Balance (now prior to default) and limit are strongly correlated at high values but only mildly correlated elsewhere (see the second plot at the bottom row of Figure 2). A similar dependence pattern is seen for full payment percentage versus paid percentage (see the sixth plot at the bottom row), 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), higher (lower) card utilisation being indicative of a lower (higher) credit score, respectively. In summary, many 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 relationships between EAD and the other variables of interest.

Since parametric copulas could wrongly specify non-monotonic dependence structures (which, as Online Appendix A points out, are observed in our EAD dataset), we extend the analysis to also include non-parametric copulas. Figure 3 displays an estimated D-vine copula 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 correlations among all 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 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

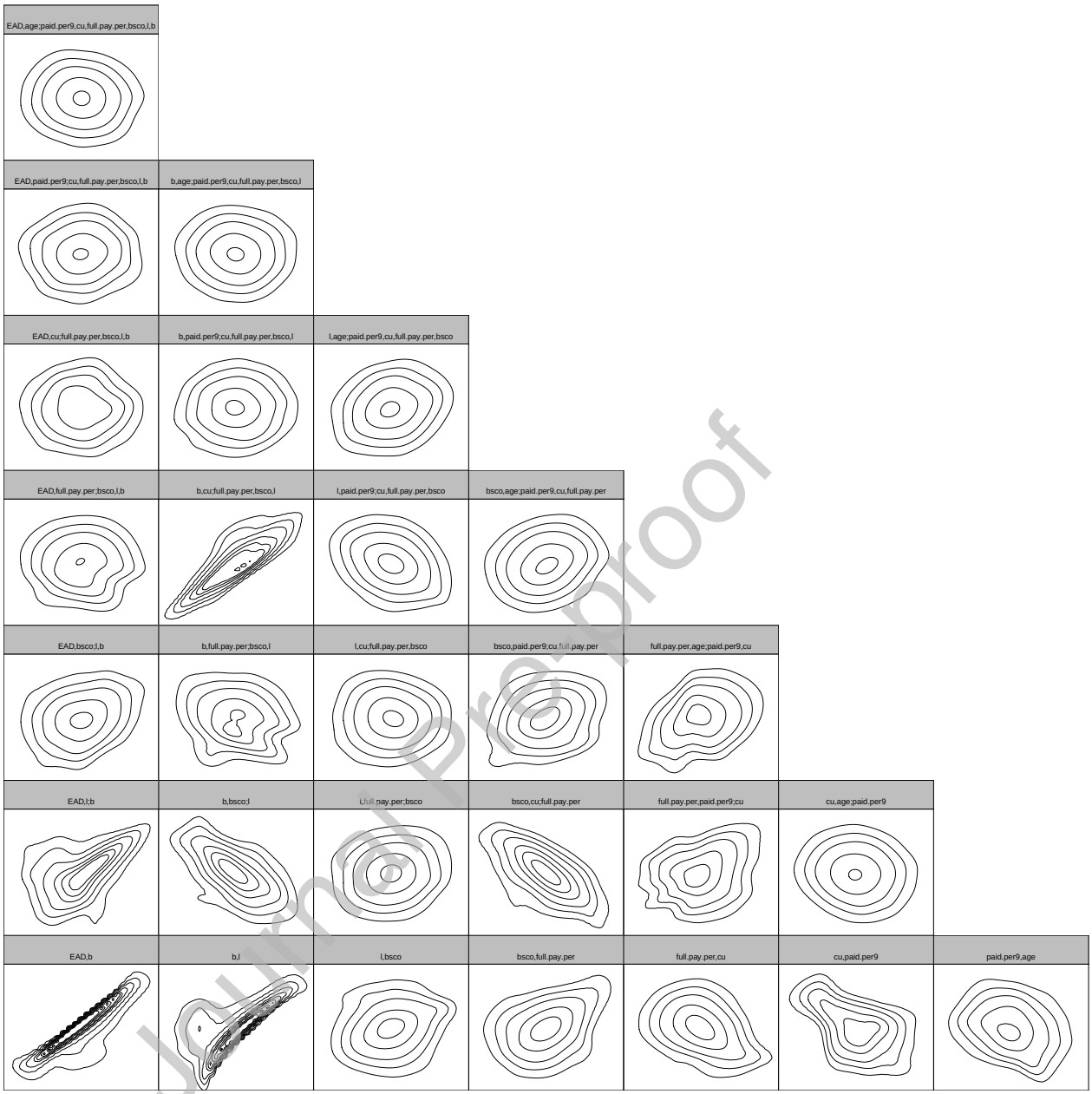
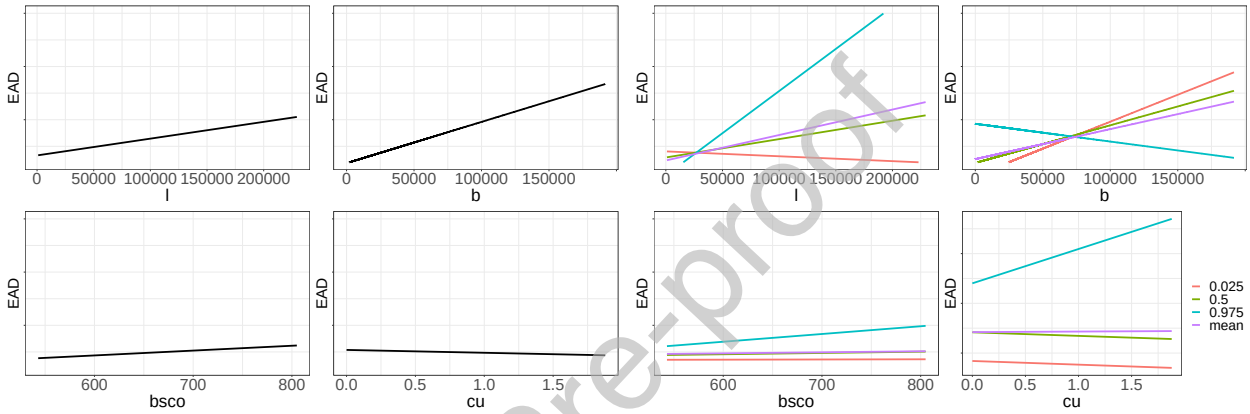


Figure 3: Estimated D-vine copula with non-parametric copulas and contour plots displaying the joint PDF of variable pairs with the first component on x-axis and the second on y-axis. The scale of both axes is (-3,3). The order of the D-vine copula is EAD → b → l → bsco → full.pay.per → cu → paid.per9 → age.

to be smaller than 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 limit.

6.2. Effects of predictors

Figure 4 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.



(a) Partial effect plots for OLS.

(b) Partial effect plots for LQR.

(c) Partial effect plots for P-DVQR.

(d) Partial effect plots for NP-DVQR.

Figure 4: Partial effect plots of a selected set of predictors on the conditional mean and 0.025, 0.5 and 0.975 conditional quantiles of EAD (with the scale of the y-axis omitted for data confidentiality reasons). The y-axes of all plots share the exact same scale.

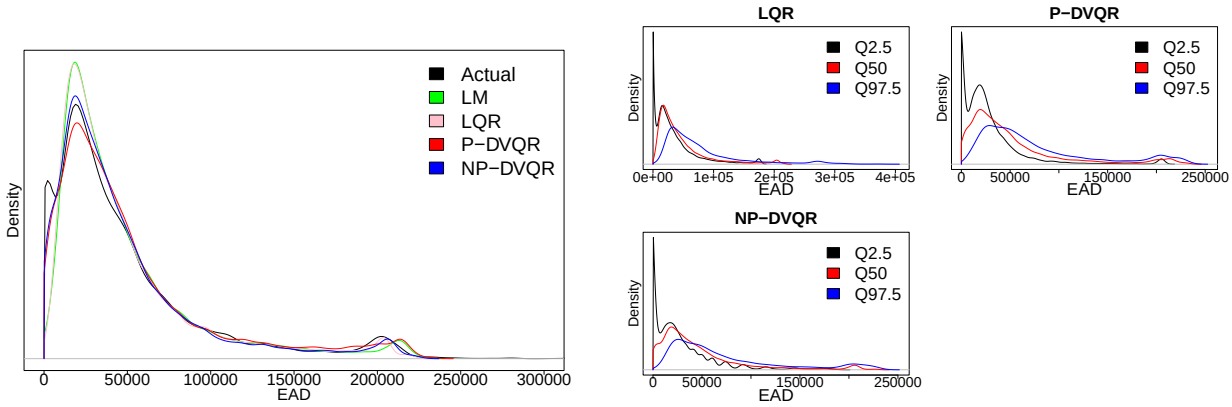
In the OLS (top-left panel), the effects on EAD mean are, by definition, all linear; considering the scale on the y-axis, balance is the variable that has the largest effect. Next, LQR (top-right panel) is able to provide deeper insights into how these effects further vary depending on the EAD

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 (and hence, for brevity, their plots are omitted).

The LQR estimates, however, are prone to quantile crossing. In the result plots 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 (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 highest credit limits (from 200,000 onward) by the bank.

6.3. EAD quantile distributions

Figure 5a 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



(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 5: Density plots of predicted EAD (with the scale of the y-axis omitted for data confidentiality reasons). The y-axes of all plots in panel (b) are all drawn on the same scale.

5b 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 upper tail, leading to the largest 97.5% value-at-risk for EAD.

6.4. 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.

6.4.1. 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.

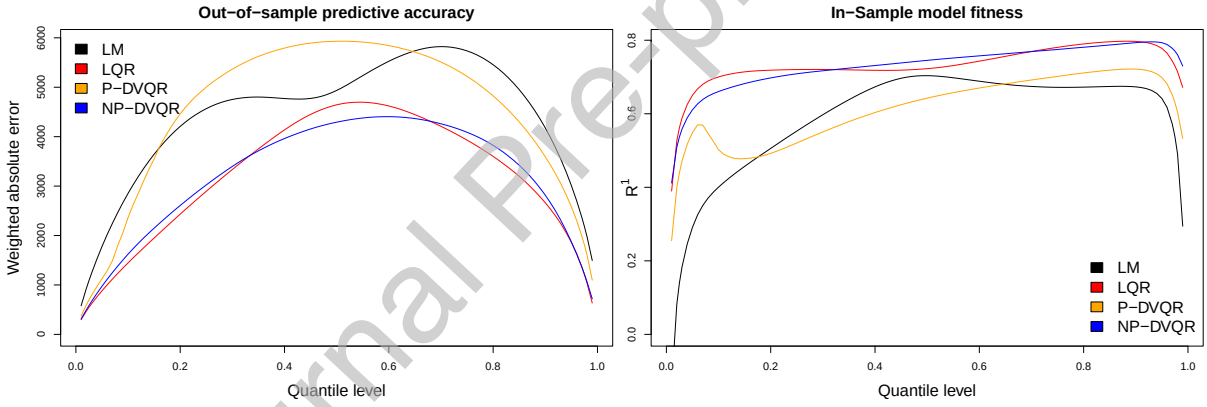


Figure 6: Performance measurements of the predicted conditional quantiles for OLS, LQR, P-DVQR and NP-DVQR: weighted absolute error (top) and model fitness (bottom).

Figure 6 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 A.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. Although being relatively close, NP-DVQR

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.

6.4.2. 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:

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

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.

Table 2 summarises the performance of all models according to these metrics.

Model	MAE ↓		LogS ↑		QS ↑		IS ↓		IBS ↓	
	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
OLS	9871	9421	-11.34	-11.21	2.19e-05	2.21e-05	62695	62764	7042	6847
LQR	9322	8863	-	-	-	-	42979	43095	-	-
P-DVQR	11400	10975	-10.48	-10.46	8.81e-05	8.91e-05	45689	46048	8677	8500
NP-DVQR	**8572**	8089	**-10.04**	-10.03	**9.85e-05**	9.65e-05	**41795**	41732	**6129**	5941

Table 2: Out-of-sample and in-sample performance results, calculated for the test set (20% of the data) and training set (80%), respectively, for point and interval estimates as well as distributions (bold face indicates best out-of-sample performance). The arrows indicate that lower values for MAE, IS and IBS, and higher values for LogS and QS, imply better performance.

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. By contrasting in-sample and out-of-sample performance of the non-parametric copula model against that of the simpler OLS and linear quantile models, we can see that its increased performance does not come at the expense of potential overfitting. Lastly, according to the interval scores, the existing heteroscedasticity can be captured by quantile models, but not OLS. For additional analyses of residual plots, we refer the reader to Online Appendix D.

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. Another interesting avenue is to extend the

vine copula-based quantile regression to LGD modeling using the two-step ahead forward selection by Tepegjozova et al. (2022) and again choose an appropriate dependence structure amongst C-vine and D-vine copulas, or even the more general R-vine copulas.

Declaration of interest

None.

Funding sources

This work was supported by the Royal Thai Government Scholarship. The authors also 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.

References

Aas, K., Czado, C., Frigessi, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics 44, 182–198. doi:10.1016/j.insmatheco.2007.02.001.

Bager, A., 2018. Ridge parameter in quantile regression models. An application in biostatistics. International Journal of Statistics and Applications 8, 72–78. doi:10.5923/j.statistics.20180802.06.

Bedford, T., Cooke, R.M., 2001. Probability density decomposition for conditionally dependent random variables modeled by vines. Annals of Mathematics and Artificial Intelligence 32, 245–268. doi:10.1023/A:1016725902970.

Bedford, T., Cooke, R.M., 2002. Vines–a new graphical model for dependent random variables. The Annals of Statistics 30, 1031–1068. doi:10.1214/aos/1031689016.

Bellotti, T., Crook, J., 2012. Loss given default models incorporating macroeconomic variables for credit cards. International Journal of Forecasting 28, 171–182. doi:10.1016/j.ijforecast.2010.08.005.

Bollen, K.A., Brand, J.E., 2010. A general panel model with random and fixed effects: A structural equations approach. Social forces 89, 1–34. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137523/, doi:10.1353/sof.2010.0072.

Bouyé, E., Salmon, M., 2009. Dynamic copula quantile regressions and tail area dynamic dependence in forex markets. The European Journal of Finance 15, 721–750. doi:10.1080/13518470902853491.

Calabrese, R., Osmetti, S.A., 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. doi:10.1111/rssa.12523.

Chang, B., Joe, H., 2019. Prediction based on conditional distributions of vine copulas. Computational Statistics and Data Analysis 139, 45–63. doi:10.1016/j.csda.2019.04.015.

Czado, C., 2019. Analyzing dependent data with vine Copulas: A practical guide with R. Lecture Notes in Statistics, Springer International Publishing. doi:10.1007/978-3-030-13785-4.

Dekking, F., Kraaikamp, C., Lopuhaä, H., Meester, L., 2005. A modern introduction to probability and statistics: Understanding why and how. Springer Texts in Statistics, Springer. doi:10.1007/1-84628-168-7.

Dette, H., Van Hecke, R., Volgushev, S., 2014. Some comments on copula-based regression. Journal of the American Statistical Association 109, 1319–1324. doi:10.1080/01621459.2014.916577.

Dissmann, J., Brechmann, E., Czado, C., Kurowicka, D., 2013. Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics and Data Analysis 59, 52–69. URL: https://www.sciencedirect.com/science/article/pii/S0167947312003131.

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. doi:10.1016/j.jkss.2015.06.002.

Geenens, G., Charpentier, A., Paindaveine, D., 2017. Probit transformation for nonparametric kernel estimation of the copula density. Bernoulli 23, 1848 – 1873. doi:10.3150/15-BEJ798.

Geidosch, M., Fischer, M., 2016. Application of vine copulas to credit portfolio risk modeling. Journal of Risk and Financial Management 9, 1–15. doi:10.3390/jrfm9020004.

Genest, C., Favre, A.C., . Everything you always wanted to know about copula modeling but were afraid to ask 12, 347–368. doi:10.1061/(ASCE)1084-0699(2007)12:4(347).

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378. doi:10.1198/016214506000001437.

Gürtler, M., Hibbeln, M.T., 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. doi:10.1016/j.jbankfin.2017.03.004.

Haff, I., Aas, K., Frigessi, A., 2010. On the simplified pair-copula construction — simply useful or too simplistic? Journal of Multivariate Analysis 101, 1296–1310. doi:10.1016/j.jmva.2009.12.001.

Haupt, H., Kagerer, K., Schnurbus, J., 2011. Cross-validating fit and predictive accuracy of nonlinear quantile regressions. Journal of Applied Statistics 38, 2939–2954. doi:10.1080/02664763.2011.573542.

Hon, P.S., Bellotti, T., 2016. Models and forecasts of credit card balance. European Journal of Operational Research 249, 498–505. doi:10.1016/j.ejor.2014.12.014.

Joe, H., 1996. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters, in: Rüschendorf, L., Schweizer, B., Taylor, M.D. (Eds.), Distributions with fixed marginals and related topics. Institute of Mathematical Statistics, Hayward, CA. volume 28 of *Lecture Notes–Monograph Series*, pp. 120–141. doi:10.1214/lnms/1215452614.

Joe, H., 1997. Multivariate models and dependence concepts. 1 ed., Chapman and Hall, London. doi:10.1201/9780367803896.

Kauermann, G., Schellhase, C., 2013. Flexible pair-copula estimation in D-vines using bivariate penalized splines. Statistics and Computing 24, 1081–1100. doi:10.1007/s11222-013-9421-5.

Killiches, M., Kraus, D., Czado, C., 2016. Examination and visualisation of the simplifying assumption for vine copulas in three dimensions. Australian and New Zealand Journal of Statistics 59, 95–117. doi:10.1111/anzs.12182.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica: Journal of the Econometric Society 46, 33–50. doi:10.2307/1913643.

Koenker, R., Machado, J.A.F., 1999. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94, 1296–1310. doi:10.1080/01621459.1999.10473882.

Komunjer, I., 2013. Chapter 17 - quantile prediction, in: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier. volume 2 of *Handbook of Economic Forecasting*, pp. 961–994. doi:10.1016/B978-0-444-62731-5.00017-8.

Kraus, D., Czado, C., 2017. D-vine copula based quantile regression. Computational Statistics and Data Analysis 110, 1–18. doi:10.1016/j.csda.2016.12.009.

Krüger, S., Oehme, T., Rösch, D., Scheule, H., 2018. A copula sample selection model for predicting multi-year LGDs and lifetime expected losses. Journal of Empirical Finance 47, 246–262. doi:10.1016/j.jempfin.2018.04.001.

Krüger, S., Rösch, D., 2017. Downturn LGD modeling using quantile regression. Journal of Banking and Finance 79, 42–56. doi:10.1016/j.jbankfin.2017.03.001.

Leow, M., Crook, J., 2016. A new mixture model for the estimation of credit card exposure at default. European Journal of Operational Research 249, 487–497. doi:10.1016/j.ejor.2015.10.001.

Loader, C., 2006. Local regression and likelihood. 1 ed., Springer Science and Business Media. doi:10.1007/b98858.

Martey, E.N., Attoh-Okine, N., 2019. Analysis of train derailment severity using vine copula quantile regression modeling. Transportation Research Part C: Emerging Technologies 105, 485–503. doi:10.1016/j.trc.2019.06.015.

Mashal, R., Zeevi, A., 2002. Beyond correlation: Extreme co-movements between financial assets. SSRN Electronic Journal doi:10.2139/ssrn.317122.

Moral, G., 2006. EAD estimates for facilities with explicit limits, in: Engelmann, B., Rauhmeier, R. (Eds.), The Basel II Risk Parameters: Estimation, Validation, and Stress Testing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 197–242. doi:10.1007/3-540-33087-9_10.

Nagler, T., Czado, C., 2016. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. Journal of Multivariate Analysis 151, 69–89. doi:10.1016/j.jmva.2016.07.003.

Nagler, T., Kraus, D., 2019. vinereg: D-Vine Quantile Regression. URL: https://cran.r-project.org/web/packages/vinereg/vinereg.pdf.

Nagler, T., Schellhase, C., Czado, C., 2017. Nonparametric estimation of simplified vine copula models: comparison of methods. Dependence Modeling 5, 99–120. doi:10.1515/demo-2017-0007.

Nelsen, R.B., 2006. An Introduction to Copulas. 2 ed., Springer New York. doi:10.1007/0-387-28678-0.

Niemierko, R., Töppel, J., 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. doi:10.1016/j.apenergy.2018.10.025.

Nikoloulopoulos, A.K., Joe, H., Li, H., 2012. Vine copulas with asymmetric tail dependence and applications to financial return data. Computational Statistics and Data Analysis 56, 3659–3673. doi:10.1016/j.csda.2010.07.016.

Parzen, E., 1962. On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33, 1065–1076. doi:10.1214/aoms/1177704472.

Qi, M., 2009. Exposure at default of unsecured credit cards. Economics working paper 2009-2. Office of the Comptroller of the Currency.

Savu, C., Trede, M., 2009. Hierarchies of Archimedean copulas. Quantitative Finance 10, 295–304. doi:10.1080/14697680902821733.

Schallhorn, N., Kraus, D., Nagler, T., Czado, C., 2017. D-vine quantile regression with discrete variables. arXiv:1705.08310.

Scheffer, M., Weiß, G.N.F., 2016. Smooth nonparametric Bernstein vine copulas. Quantitative Finance 17, 139–156. doi:10.1080/14697688.2016.1185141.

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., Whittaker, J., 2007. Quantile regression for modelling distributions of profit and loss. European Journal of Operational Research 183, 1477–1487. doi:10.1016/j.ejor.2006.08.063.

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F., 2017. Flexible regression and smoothing: Using GAMLSS in R. Chapman and Hall. doi:10.1201/b21973.

Stöber, J., Joe, H., Czado, C., 2013. Simplified pair copula constructions—limitations and extensions. Journal of Multivariate Analysis 119, 101–118. doi:10.1016/j.jmva.2013.04.014.

Tepegjozova, M., Zhou, J., Claeskens, G., Czado, C., 2022. Nonparametric C- and D-vine-based quantile regression. Dependence Modeling 10, 1–21. doi:10.1515/demo-2022-0100.

Thackham, M., 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, 1267–1286. doi:10.1111/rssa.12418.

Tong, E.N., Mues, C., Brown, I., Thomas, L.C., 2016. Exposure at default models with and without the credit conversion factor. European Journal of Operational Research 252, 910–920. doi:10.1016/j.ejor.2016.01.054.

Valvonis, V., 2008. Estimating EAD for retail exposures for Basel II purposes. The Journal of Credit Risk 4, 79–109. doi:10.21314/jcr.2008.069.

Van Gestel, T., Baesens, B., Van Dijcke, P., Garcia, J., Suykens, J.A., Vanthienen, J., 2006. A process model to develop an internal rating system: Sovereign credit ratings. Decision Support Systems 42, 1131–1151. doi:10.1016/j.dss.2005.10.001.

Wattanawongwan, S., Mues, C., Okhrati, R., Choudhry, T., So, M.C., 2023. A mixture model for credit card exposure at default using the gamlss framework. International Journal of Forecasting 39, 503–518. doi:10.1016/j.ijforecast.2021.12.014.

Yu, R., Yang, R., Zhang, C., Špoljar, M., Kuczyńska-Kippen, N., Sang, G., 2020. A vine copula-based modeling for identification of multivariate water pollution risk in an interconnected river system network. Water 12. doi:10.3390/w12102741.

Zhi, B., Wang, X., Xu, F., 2020. Impawn rate optimisation in inventory financing: A canonical vine copula-based approach. International Journal of Production Economics 227, 107659. doi:10.1016/j.ijpe.2020.107659.

Zhu, K., Kurowicka, D., Nane, G.F., 2021. Simplified R-vine based forward regression. Computational Statistics and Data Analysis 155, 107091. doi:10.1016/j.csda.2020.107091.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: