Credit Risk in FinTech Organisations: a Copula Approach for Default Dependency

Matteo Comelli*

August 28, 2019

Abstract

FinTech businesses, in particular P2P lending platforms are currently under a fast paced growth in Europe, America and Asia. This report will analyse the main credit risk characteristics with the use of copulas for one of the major player in P2P lending platforms (Lending Club), with attention to default dependencies of the loans issued. The key aspects in which this report focus on are the main drivers for credit risk: credit scoring, default dependencies and financial strategy tied with the risk exposure that is the core of P2P lending platforms business. What we have found was that old methods like Gaussian copula are obsolete and that often segment grading from FinTechs is not appropriate with the market standards where disparities arise together with macroeconomic issues highlighting the difficulties in grade selection. The use of a high timespan of 8 years permits to consolidate our results and the solid presence of Lending Club in the market gives us the potentiality to infer possible extreme events dynamics within established FinTechs.

Keywords: Copulas, Default Dependency, P2P Lending Platform, FinTech, Credit Risk, Lending Club.

1 Introduction

The objective of this report focuses on the dynamics of "peer-to-peer" (P2P) loan lending platforms with the analysis of default dependencies and credit risk. In particular, the Lending Club (LC) loans data are used for building credit risk models with copula methods.

This report will contribute to the credit risk management and Fintech literature by exploring default dependencies of loans. We use the classic copula methods most commonly used in credit risk within the banking sector and apply it to the new contest of P2P lending platforms. The main contribution of the report is that it gives an initial approach to regulators, new FinTech businesses and academics to establish and gauge a better credit grading oversight in the FinTech industry.

Copula methods are widely used for risk management however, the past risk management literature

*University of Sussex Business School. Email: mc696@sussex.ac.uk

has not always applied copulas in the most effective way. If not appropriately tested in the industry they may lead to controversial results and biased estimations. Some blame a wrong approach in copula application for risk measurement with the Gaussian model particularly in investment banking as one of the main causes for the financial crisis in 2008-2009 (14). For the P2P lending platforms not much attention is devoted to how and what methods are applied for risk modelling, where the necessity for more regulation could be seen as necessary for improving operational and credit risks that are significantly high in growing and new businesses.[1]

The failure of past Gaussian copula models is documented in the literature (see Section 1) and we study how this apply for the credit risk evaluation in the Fintech industry with focus on loan default dependencies of Lending Club data.

Risk and Reward in P2P Lending Platforms

P2P lending platforms detach themselves from bank institutions business model by few key characteristics: they do not take deposits from the borrowers neither lend themselves directly with loan contracts.

Avoiding the deposits permits to P2P lending platforms to take no risk in their balance sheet. Furthermore, the contracts are issued by a partner bank and not the P2P platform (see figure 1) which issues only notes to the clients (8).

The main source of income for P2P is from fees and commissions generated by providing the service of linking borrowers and investors looking for high returns.

This report will focus more on the risks behind P2P lending platforms related to: loan acceptance rate often tied with macroeconomic events (see section 3), the grading of segments (mainly generated internally with the possibility of biased results), the probability of dependency of defaults when diversification between obligors cannot hedge that risk and finally the absence of collaterals that exposes the investors directly to default risk.

P2P Lending Platforms: The Business Model

The core of the risk is moved by the P2P lending platforms to the investors accordingly to the business model adopted (see figure 1).[2]

[1] FinTech are not graded by external Credit Rating Agencies

[2] (8)

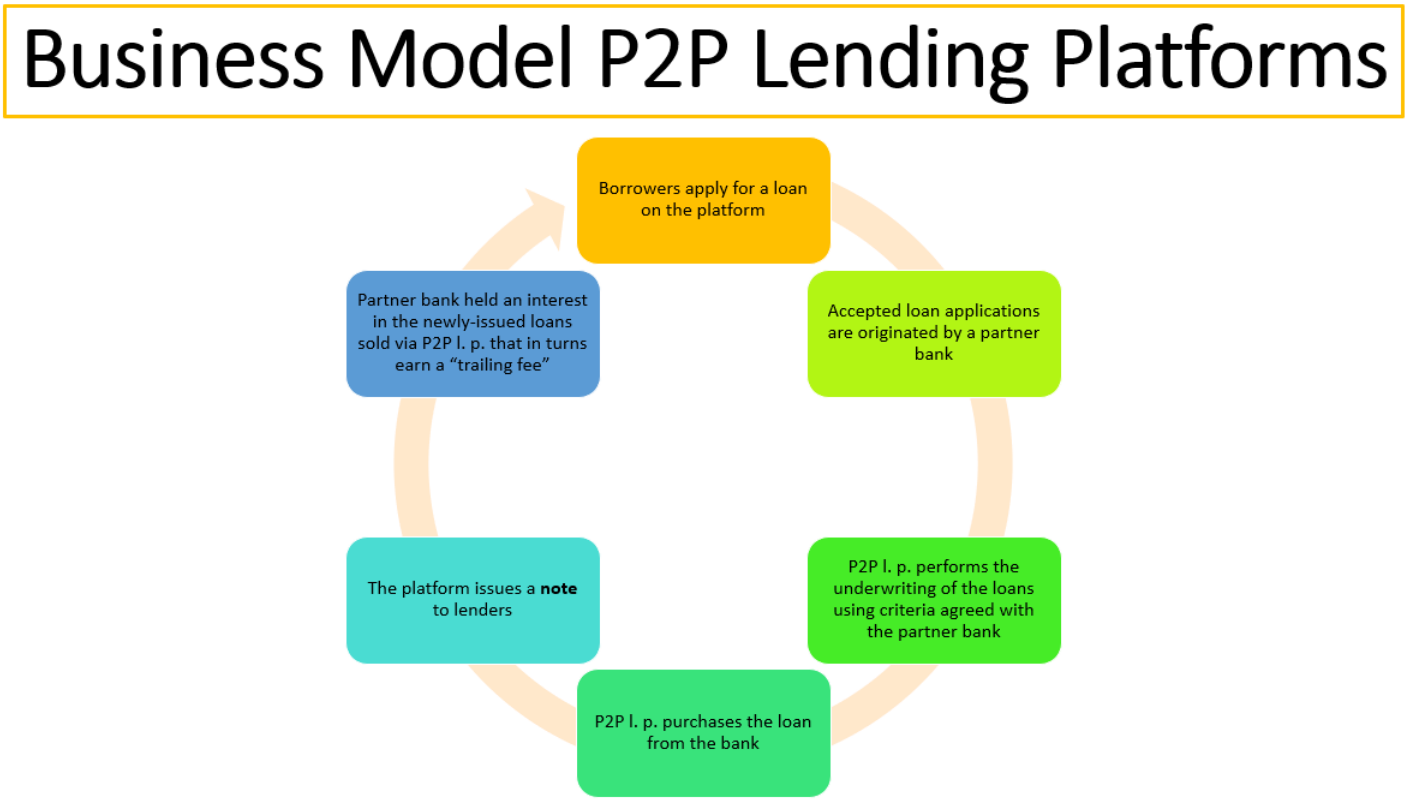


Figure 1: LC Business Model Outlook

The main advantage of this business model is that it works as a complementary market of banks by fulfilling the needs of credit of high-risk or low-income borrowers that banks will never accept. P2P lending platforms differentiate themselves also by applying innovative risk scoring and digital channels as example the Lendoit blockchain based platform.[3]

P2P Lending Platforms: The Impact of Regulation

The regulatory role is gaining importance since the well established FinTech businesses are facing credit crisis.[4] Differently from bank institutions P2P lending platforms had little or not regulation, where major concerns involve money protection rules (since P2P are not covered by the government FSCS for the first £75000 of deposit),[5] or conduct where the P2P lending platforms assign the risk grade score to segments by themselves with the risk of inaccurate scoring level. Recently, one of the major European FinTech Funding Circle tightened its lending criteria because of slowing down of demand and rise in bad debts in order to align its standards with the market.[6]

Case Study: Lending Club

Lending Club is market leader in personal loans, its business model can be generalized for all FinTech organizations and it can be defined as the benchmark for P2P lending platforms.

[3]https://medium.com/@chima825/lendoit-revolutionizing-the-future-of-marketplace-lending-ebad8cb3bc44
[4]https://www.ft.com/content/d44b23e0-9c90-11e9-9c06-a4640c9feebb
[5]Financial Services Compensation Scheme
[6]https://www.ft.com/content/cd13abc0-b9a9-11e9-96bd-8e884d3ea203

The decision to use Lending Club dataset is based on the fact that it can be seen as the major player in FinTech industry with $35 trillion issued over the past 11 years. Lending Club was founded in USA (California) since 2007 counting now more than 3 million customers and more than 200,000 investors.[7]

For the Lending Club case we observe 11 years (2007-2019Q1) of data to construct our variables during which it suffered several crisis however, the final time frame considered does not include the period 2007-2010 with the financial crisis drastic effect. The omission of this period is not crucial for a complete risk analysis since Lending Club suffered of financial and operational crisis during the final dataset considered (2010-2018) thus, we consider possible unexpected extreme events and how those affect the default dependencies. The dataset considered is significantly longer with 8 years of data. It permits to explore the credit risk in P2P lending platforms with a more grounded and solid back up of data for our assumptions compared to past literature.

Report Overview

Following the introductory Section 1 the literature review (Section 2) reassesses the main literature in credit risk management and FinTech by exploring default dependencies modelling. The initial data elaboration is described in Section 3, where we compute default and relative loss (RL) rates from LC dataset. Section 4 is a theoretical review of copula functions, properties and limitations. Section 5 outlines the methodology steps (4): 1-Kernel density estimation (KDE),[8] 2-Copula densities calculations, 3-Copula fitting process and 4-Copula selection technique.[9]

Section 6 gives the main results from the copula density comparison; finally, Section 7 concludes.

2 Literature Review

Default dependency models are critical for assessing credit risk, where the position of peer to peer lending platforms in risky loans need more analysis centred on this aspect for default dependencies. Economic crisis (2008-2009) showed how extreme events are happening more frequently than excepted and that traditional models were not being able to predict or at best react promptly to those shocks. Moreover, the novelty of FinTech businesses is tied with initial low regulation and high possibility of complications compared to other competitors more established in the lending market like bank institutions. The underestimation of risks was spread in the literature according to Fenech (10) for loan-loss tail dependencies. By the end of 2009 the Gaussian copula approach advocated by Li (13) was not appropriate any more (10).

While the essence of a good credit analysis can be summarized by the ability in determining repayment probabilities (1) the default dependencies are lately taking over the attention of the public for credit risk in particular for those businesses highly affected by credit grading as FinTech industries. In particular for P2P lending platforms, as they differ from traditional lending of bank institutions,

[7] https://ir.lendingclub.com/CorporateProfile

[8] We apply to the variables the kernel density estimation in order to work with the uniform variables and copula density functions.

[9] Goodness-of-Fit

loans are defined as high risky and if rated through credit reporting agency the majority would be graded as high risk ones where default events are labelled with "high" or "very high" risk level (19)[10].

Different asset classes for the loan case cannot be considered a certain method for minimising the portfolio risk by diversification when we consider systematic default dependencies. Particular asset classes might be correlated and show dependence structure in defaulting scenarios (10) and for P2P lending platforms the major concern is to capture this systematic dependency.

Loan default dependencies have relevant implications for institutions considering their portfolio risk tolerance. Lending Club in particular has been accused of funding loans with its own investment basket enhancing in turn the risk-free exposure that it advocates within its loans[11].

Main assumptions in traditional models from the past literature are characterized by a linear relationship between extreme events. The main difficulties arise when modelling dependent risk factors is not directly represented by multivariate normal distributions however, in this scenario copulas become the natural application for risk dependencies and of useful use (15). Joint distributions give us the idea of how different components move together where copulas help to understand this dependence, copulas can be seen as an analytical tool when correlation is not good as measure of dependence.

Copula approach was first introduced by Sklar in 1959 (18) as statistical mechanism to transfer the joint distribution into its marginals to show the dependence structure. Several applications of copulas into finance followed: from option pricing with copula approach (5) to first applications to risk by Li (13) in tail dependency of default corporate correlations and for choices more related to credit risk management of portfolios with Crook and Moreira (7) and before by Dematra and McNeil (9). Copulas approach became popular because of model building: from knowing a lot about the marginal distributions but little about the joint distributions, thus building a model for simulations that could respect the marginal relations was tempting.

Risk management theories are mainly built on multivariate normal i.i.d. return distributions, result not empirically true (15). Copulas instead are flexible tools for creating joint distributions (3), they take apart the dependence structure from the marginal distributions, indeed with same marginals different copulas can produce different joint distributions. Since it is possible to find the same Pearson correlation value for two joint distributions from different copulas, the correlation coefficient cannot capture the risk characteristics of the distributions because built as symmetric linear dependence metric (3). Linear correlation depends on the marginal distributions, thus is invariant under strictly increasing linear transformations, on the other hand rank correlation coefficients are always defined and invariant with strictly increasing transformations following the invariance principle (15).

The holistic approach where Gaussian copula is able with any margins to produce sufficiently many extremes values for example on the tails is obsolete and dangerous as seen in 2008 (15). Moreover,

[10]https://towardsdatascience.com/p2p-lending-platform-data-analysis-exploratory-data-analysis-in-r-part-1-32eb3f41ab16

[11]https://www.businessinsider.com/lendingclub-faces-doj-and-sec-investigations-could-buy-more-of-its-loans-2016-5?r=USIR=T

Gaussian Copulas failed to detect the correlations of loan default across various asset classes (10). Other types of copulas are instead more appropriate as they capture asymmetric loan default distributions (10), as expected risk managers are more and more keen on capturing tail dependence of defaulting loans.

Studies have confirmed the non-normality hypothesis of loan portfolios (Rosenberg and Schuermann 2006: in (7)) also supporting the asymmetric tail dependence (Das and Geng 2006 in (7)) where Di Clemente and Romano (2004 in: (7)) presented the case for returns of credit assets. Subsequently the default rate for credit loans tend to increase during recession and lower in booming periods, implying that between each credit there is a possible form of dependence (13) that can be captured with copulas. The essential step is to detect the systematic default dependence during those periods between obligors rather than individual loans dynamics that can be marginally avoided by diversification in the loan issue process for P2P lending platforms.

Latest models are taking into consideration the asymmetric behaviour in credit asset returns and the non-normality of the distributions (7) with copulas from the Archimedean family. The success claimed for the t-copulas presented by Dematra and McNeil (9) is concrete as t-copula have the ability to capture the phenomenon of dependence of extreme values significantly better than Gaussian methods. However, for the P2P loan lending platform case asymmetric copulas are used to create model of default dependency that properly fit the default data. For Lending Club data the Archimedean family of copulas permits to adapt the copula to extreme event in the tails in joint default events.

In Section 4 we will present the main copulas applied and in Section 6 the results of this analysis.

3 Data Elaboration

The empirical study is based on loans data from the Lending Club dataset,[12] that is considered as the "gold standard" of peer-to-peer lending platforms.[13]

Lending Club was founded in 2007 in California (USA) and it was forerunner of the idea of reinventing credit and investing within FinTech industries having a clear success during the past 11 years.

The dataset contains all loans issued with a monthly frequency for the 12 years from 2007 to 2019 until the first quarter, where each loan is classified with its credit segment grade quality by Lending Club. Segment A represents the least risky loans, while G the riskiest. Other segments listed are: B, C, D, E and F.

Credit risk literature presents various methodologies for computing default rates. In one paper where Crook and Moreira (7) analysed credit card portfolio asymmetric dependence they computed the default rates as the amount of loans that reached 90 days in arrears divided by the number of active accounts in the same month. Similarly, we computed the default rates with the formula

[12] https://www.lendingclub.com/info/download-data.action

[13] https://www.businessinsider.com/lendingclub-faces-doj-and-sec-investigations-could-buy-more-of-its-loans-2016-5?r=USIR=T

derived from the amount of loans defaulted and not-defaulted as:

$$DR = \frac{x_{jt}}{n_{jt}} \tag{1}$$

Where j ranges between the grades of the segments (A, ..., G), t stands for the specific month thus t=(1, ..., 104) for a total of 104 months (2010-2018), x represents the number of defaulted loans and n is the total amount of defaulted and non-defaulted loans.
Nonetheless, we are more interested into the patterns of the default dependencies in relative losses (RL) where defaulted amount will be net by the recovery rate. This approach will focus more on the amount of loans rather than volume. In the FinTech industry the volume issue affects loan acceptance rate and LC dynamics might influence the default rates computed from EQ(1) more than relative loss rates (see figure 5). For supporting our statement we plot figure 2 to see the relative loss rates, by taking a closer look at the period 2014-2018 in figure 3 where LC suffered more. All the RL rates jumps by the end of 2014 and start to have a wider trend after May 2016 by increasing the gap between the segments. For RL rates we see less sensitivity than in the volume of loans where movements are huge: figure 5 where the dotted black line shows 2014 IPO and 2016 CEO resignation.

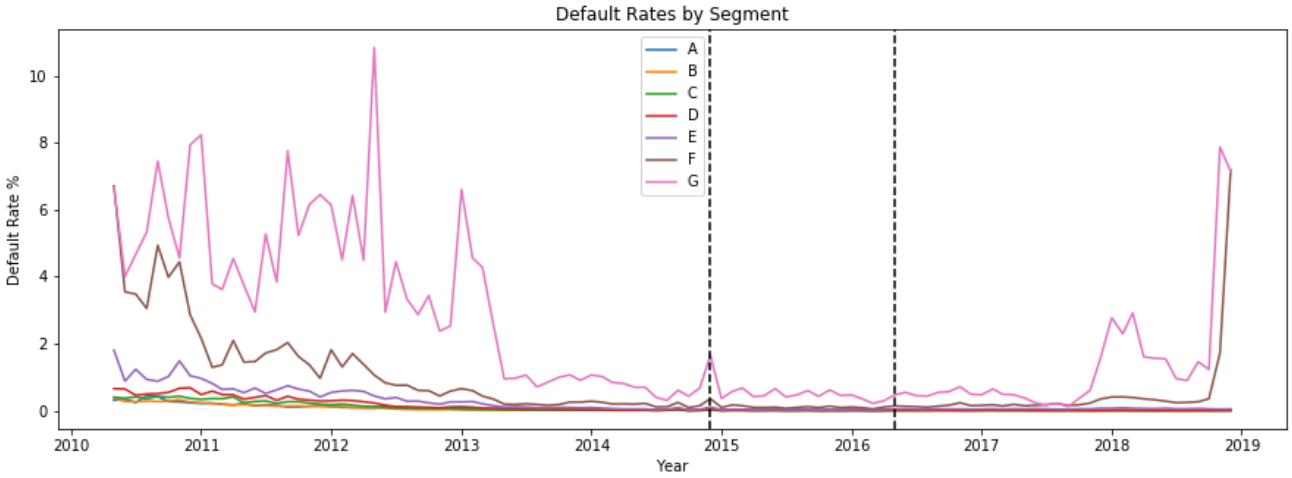


Figure 2: Relative loss rates per segment, the black dotted line shows the unusual events affecting RL rates.

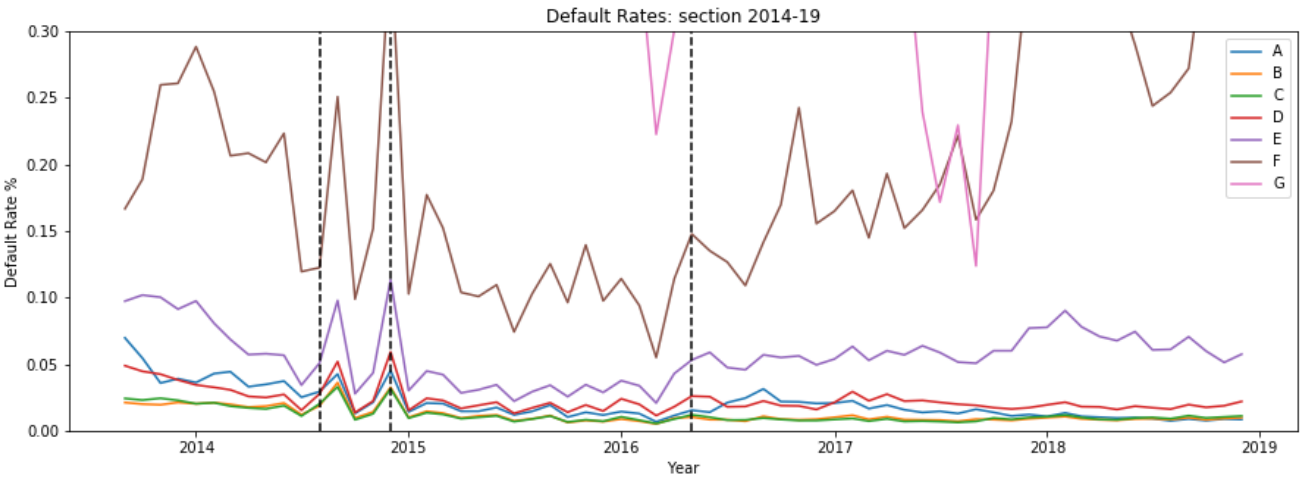


Figure 3: Relative loss rates per segment: 2014-2018.

Considering this approach as starting point then we computed the relative losses for each segment. First selection takes into consideration the loan status: defined as defaulted all loans labelled as "Charge-off", "Delay from 60 to 120 days" and "Defaulted". On the other hand, the loans not taken into consideration for the default rates computation were those marked as "Fully Paid" and "Current".

The relative loss rates (RL) are computed from EQ(2):

$$rl_{it}^{j} = \frac{(l_{it}^{j} - r_{it}^{j})}{m_{it}^{j}} \tag{2}$$

Where rl represents the relative loss rate for j segments and t months (up to 104), i is the single loan amount observed from the dataset. For the second argument of the equation l is the loan amount, r the recovery rate and finally, m as total amount borrowed in segment j.

The relative loss rates were computed by segment on the monthly base, this means that all relative loss rates according to our selection criteria were averaged together for each segment and month. As result we obtained 132 relative loss rates per segment for the time frame 2007-2018 of 11 years thus 924 monthly observations.

The final dataset used for relative loss rates analysis does not take into account the first three years: from January 2007 to April of 2010 leaving the final dataset with a total of 104 months from May 2010 to December 2018. This decision is made because of the low amount of loans during the first years (2007-2009) when Lending Club started to operate. The major argument is that the sensitivity to volume for relative loss rates is significantly affecting the results, hence deleting periods with low amount of loans issued is justified. The first four months of the year 2010 were deleted because were following the argument of the first three years of operation for Lending Club, indeed they were showing significantly higher relative loss rates, thus considered as outliers in a first screening. In conclusion, the time frame has been selected when Lending Club was well established in the market in order to not be misguided by the novelty of the Fintech business and abnormal relative loss

rates. The period considered does not include the 2008–2009 years where the market was suffering a significant financial crisis however, the low number of loans would have given us biased values for relative loss rates. On the other hand, Lending Club passed through several scandals and issues within investment decisions that are in turn represented in the loans acceptance dynamics.

Figure 4 presents the pairwise representation of the loans per segment for the final dataset.

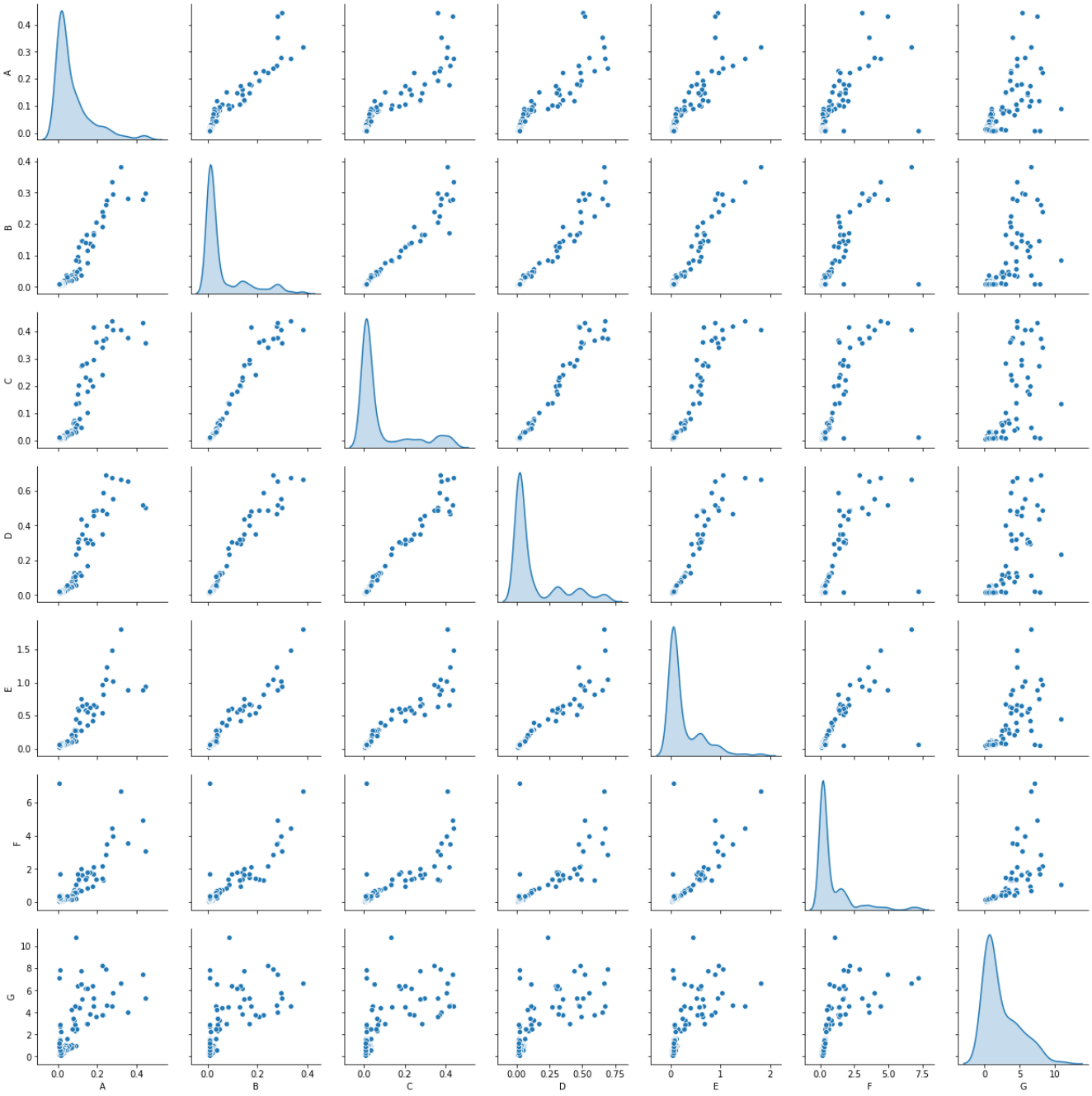


Figure 4: Pair plot representation of RL rates of all segments A-G, x and y axis shows the RL rates in percentage terms, the diagonal shows the distribution estimation for each segment.

The distributions are shown on the diagonal in figure 4, we can immediately see that all the segments

present a significant positive skewed distribution, even if few segments (B, C, D) present more than one peak those cannot be defined as multi-modal distributions. Moreover, all segment distributions reject the null hypothesis of normality from the Jarque-Bera test (11) with a possible exception of the segment G that seems tending to the Guassian distribution.[14] The segment G seems slightly different from the others for its right fat tail that does not decrease as rapidly as for the other segments. For this segment we have to consider again the volume as a key element affecting the distributions, the amount of loans for the segment G is lower than any segment in our dataset on average.

Next figure 5 shows the number of loans for all the segments: the blue line depicts the number of not defaulted loans whereas the red line represents the number of defaulted loans. The black dotted lines stands for the unusual events that hit LC during its operation (IPO and CEO resignation).

The segment G is the lowest for both amount of loans defaulted and total. In figure 5 the second lowest segment is F that is still higher than G for both measures on average but significantly lower than the other segments by peaking only at 1700 loans, for comparison segment A peaks at roughly 15000 loans. We can see how LC relies on the first segments (A, B and C) for its business and how the unexpected events affect significantly all the segments, even the most secure ones like segment A and B.

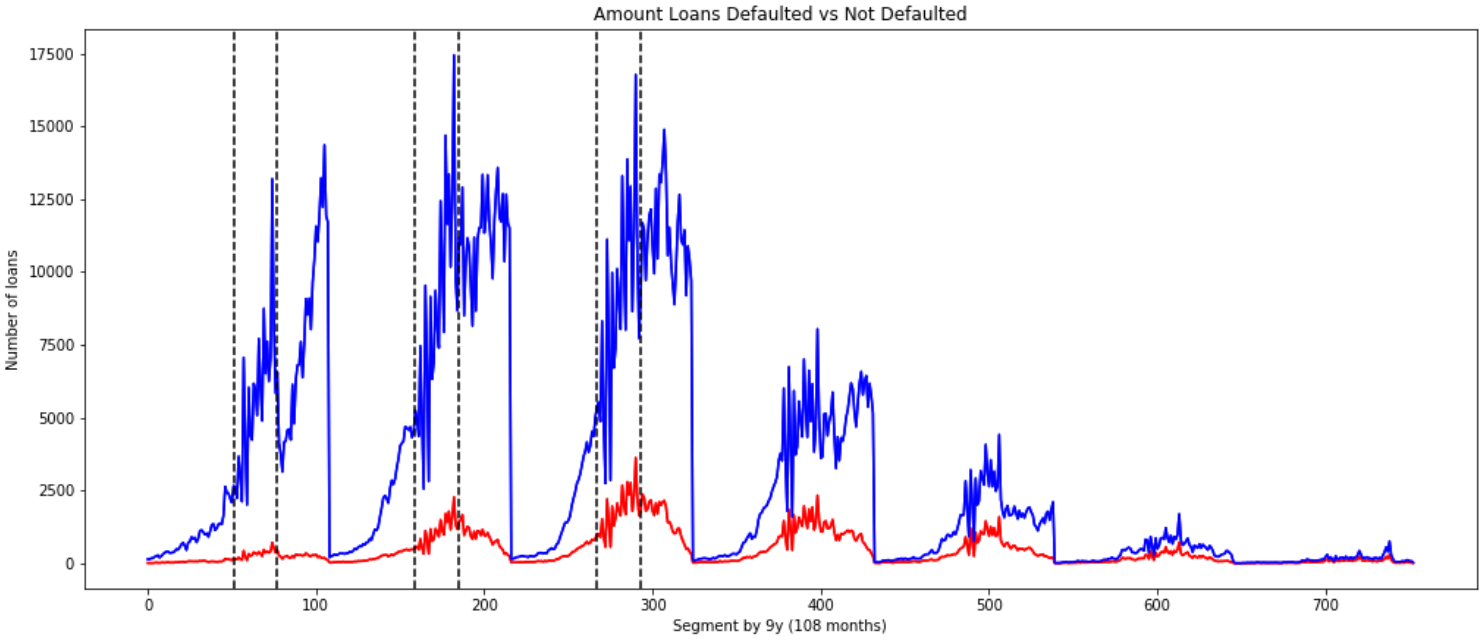


Figure 5: Number of loans for each segment, each wave represents 9 years of data of a different segment starting by grade from A to G.

Figure 6 shows the correlation of the RL rates between the segments over the years.

[14] CI depending on the confidence interval chosen α

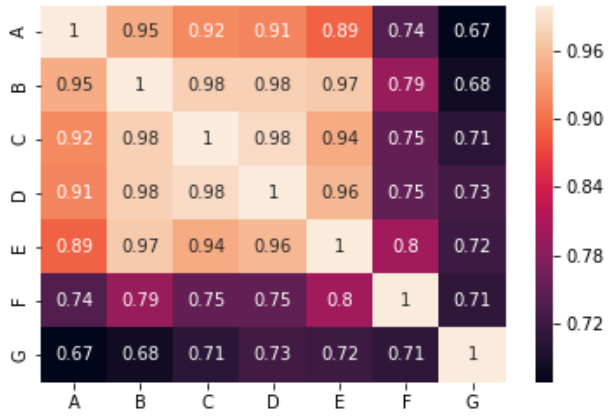


Figure 6: Correlation matrix of all segments.

As we expected the correlation between low risk segments decreases with high risk segments. Table 1 shows the summary mean for each segment:

Table 1: 1: Relative Loss Rates Summary Stats - Mean

A	**B**	**C**	**D**	**E**	**F**	**G**
0.076%	0.061%	0.088%	0.135%	0.260%	0.848%	2.376%

Except segment A that has a slightly higher relative loss rate mean than B, all the other segment rates increase accordingly with the risk level. From segment D to G the RL rate jumps significantly, this fact probably shows that the segment grade selection for loans of Lending Club can be improved for a better interpretation of the gap between segments D, E and F.

The next step is applying the Kernel Density Estimation (KDE) to the margins, to better understand the relative loss rates and obtaining the uniform transformation. KDE permits to fit the data to a probability distribution function from where we will be able to create the uniform variables, for example: picture 7 shows how the frequency histogram of the segment A relative loss rates and how is represented by the non-parametric univariate KDE in figure 8.

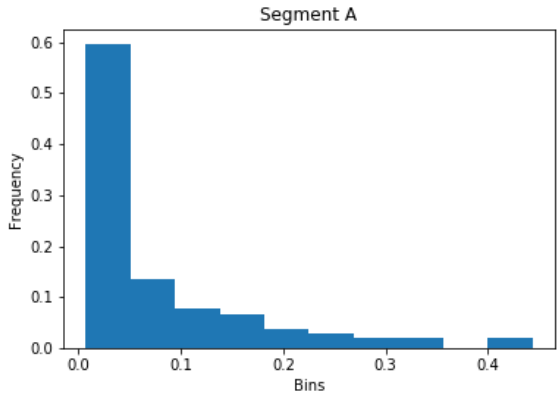


Figure 7: Histogram of Segment A with frequencies.

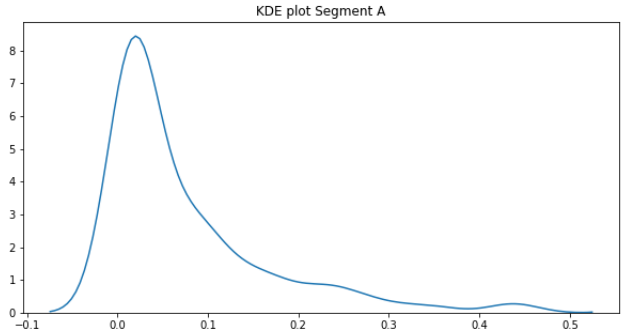


Figure 8: Segment A KDE distribution.

The KDE result is essential in determining the cumulative distribution function (CDF) for each segment from where the parametrisation of each copula can take place by using canonical maximum likelihood (CML) from the uniform variables (see Section 5). We estimated from the CDF of the KDE the variables within Python 3 environment; see Section 5 for the methodology used,[15] figure 9 for the results.

4 Copula Theory

The post 2008 events triggered concerns with respect to the Gaussian approach with default dependency models. Modelling loan default correlation was based on normally distributed data and linearity in correlation that did not emulate appropriately the market. The use of a more sophisticated and exhaustive methodology centred academics and businesses to focus on the copula approach for risk management measurement.

Copula Properties

[15] see methodology section for libraries and KDE applied

Copulas were first published in the 60s with Sklar (18) and Li (13) was the first that applied copulas in credit risk. The most frequent copulas considered were the elliptical copulas such as Gaussian and Student-t and from the Archimedean family the Clayton, Frank and Gumbel copulas. In this analysis we will focus mainly on the Archimedean copulas because of their asymmetric properties in capturing extreme default dependencies in the tails.
Copulas permit to see potential pitfalls of approaches to dependence that focus on correlation. Overall copulas have the ability to provide useful modelling techniques where bivariate distribution functions with continuous margins may be uniquely represented (10).

Sklar's Theorem

Specified by Sklar, It shows that all multivariate distribution functions contain copulas and that copulas can be used with univariate distribution functions to construct multivariate distributions (18).
Any combination of margins can be selected to build the joint aggregate so through copulas risk managers facilitate the construction of joint distribution functions. Moreover, any joint distribution can be split up into a section where information related to the variables are clear and a part which captures the dependency structure from the joint distribution function (15). Sklar's theorem permits to move from the joint distribution function to the unique copula EQ(3) and conversely as seen in EQ(4):

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)). \tag{3}$$

$$C(F_1(x_1), F_2(x_2)) = F_1(x_1)F_2(x_2). \tag{4}$$

Sklar's theorem shows that given the marginal distributions, distinct copulas define specific joint densities (3). Hence, given $F(x_1, ..., x_n)$, with continuous marginals the unique copula C will be defined by EQ(5):

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \tag{5}$$

Gaussian Copula Properties and Limitations

The most important implicit copula is the Gaussian copula, it has a correlation matrix for parameters and for this convenience it has been largely used in risk management (3).
It is derived from the multivariate and univariate standard normal distribution functions (Φ) and defined by:

$$C(u_1, ..., u_n; \sum) = \Phi(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)). \tag{6}$$

Gaussian copula's correlation coefficients depend on the copula function and also on their marginal distributions thus, the measure of dependency varies with the changes in the scale of marginal variables (10). For example standardization as strictly increasing transformation of the marginals when working with a Gaussian do not affect the copula thanks to the invariance principle (10) however, linear correlation coefficients are affected by strictly non-linear increasing transformations.

The Gaussian copulas advantages are the ability to modelling pairwise dependencies, the availability of the density function and simplicity in sampling. On the other hand, main shortcomings of Gaussian copula can be listed as: inability to model the dependency in tails because radially symmetric thus, it assigns low probability to the extreme events, for the P2P lending platforms case: large loan losses. Assumption of normality of the underlying random variables together with the linear nature of correlation between credit events are the two main limitations associated with Gaussian copulas within a loan default correlation context.

The Archimedean Family

For a non-normality approach to default modelling the Archimedean family is appropriate to capture the dependency in tails more accurately than Gaussian.

Archimedean are parameterized by their specific generator function denoted by the symbol Ψ(u), differently from Gaussian which is based on the normal distributions. Correlation structure for Archimedean is related to Kendall's tau. The tau measures the monotonicity between variables and it is invariant under strictly increasing transformations of the underlying random variables (15). The tau is only dependent on copula, whereas the linear correlation in Gaussian copula is variant under strictly increasing transformations of the underlying variables. The greatest advantage of this family is that Archimedean measures dependency solely on the copula itself.

As result Archimedean copulas are more accurate and suitable for credit risk modelling. The explicit way of constructing the copulas in terms of the generator function Ψ(u) used facilitate the application of this family. Moreover, the not restricted radial symmetry gives the possibility to capture the extreme default events in the tails.

Gumbel Copula

From the generator function:

$$\Psi(u) = -(ln(u))^{\delta}, \delta \geq 1. \tag{7}$$

For the Gumbel copula density in two dimensions:

$$C(u_1, u_2; \delta) = (A + \delta - 1)A^{1-2\delta}exp(-A)(u_1, u_2)^{-1}(-lnu_1)^{\delta-1}(-lnu_2)^{\delta-1}. \tag{8}$$

where $A = ((-lnu_1)^{\delta} + (-lnu_2)^{\delta})^{\frac{1}{\delta}}$.

It is non-linear and captures extreme events in the tail. Its asymmetry permits to control right upper tail dependence (10) with weak lower tail dependence. When delta equals 1 then it shows independence.

Clayton Copula

From the generator function:

$$\Psi(u) = \alpha^{-1}(u^{-\alpha} - 1), \alpha \neq 0. \tag{9}$$

The Copula is defined:

$$C(u_1, ..., u_n; \alpha) = (u_1^{-\alpha} + ... + u_n^{-\alpha} - n + 1)^{-\frac{1}{\alpha}}. \tag{10}$$

Introduced by Clayton (6) it has asymmetric tail dependence and mainly used to detect lower tail dependence however, it lacks the upper tail one. Lower tail dependence increases as the degree of dependence increases (10).

5 Methodology

In this section the creation and selection of copulas is carried out by giving an outlook of the whole procedure. Firstly, the kernel density estimation (Step 1) uses the variables generated in section 3 to infer the cumulative density distributions (see figure 9). Secondly, copula density functions (Step 2) are computed for all segments' pairs. Finally, the copula parametrisation section (Step 3) permits to fit the copula to our data where the selection methodology (Step 4) highlights the most appropriate one.

Step 1: Kernel Density Estimation

Non-parametric models have the peculiarity of not being specified a priori but determined from the data. Among those models, kernel density estimation provides estimates better representative than histograms for creating the probability density function (PDF) of a random variable from a random vector. In particular, the method used can be found as Python 3 command.[16]

The purpose of kernel fitting is to derive a smooth line from a random vector to represent the optimal probability density of the variable considered. It infers the population density from an empirical density function (2) (see figure 8 as example).

Kernel functions depends particularly on the bandwidth that represents the width of the histogram. The kernel finds the optimal bandwidth in order to minimize the asymptotic mean integrated square error (MISE)(2).

One common reference for bandwidth is Silverman's rule or normal distribution approximation (17) that we applied in Python non-parametric univariate KDE with Gaussian kernel.

Figure 9 gives us the cumulative distribution functions derived by KDE for all credit segments of our dataset, the blue dotted line represents the normal distribution CDF for comparison while the black line is the closest to normal representing the segment G. This unusual result permits to see how the riskier segment has a bigger upper tail than all the others.

[16] http://www.statsmodels.org/devel/generated/statsmodels.nonparametric.kde.KDEUnivariate.html

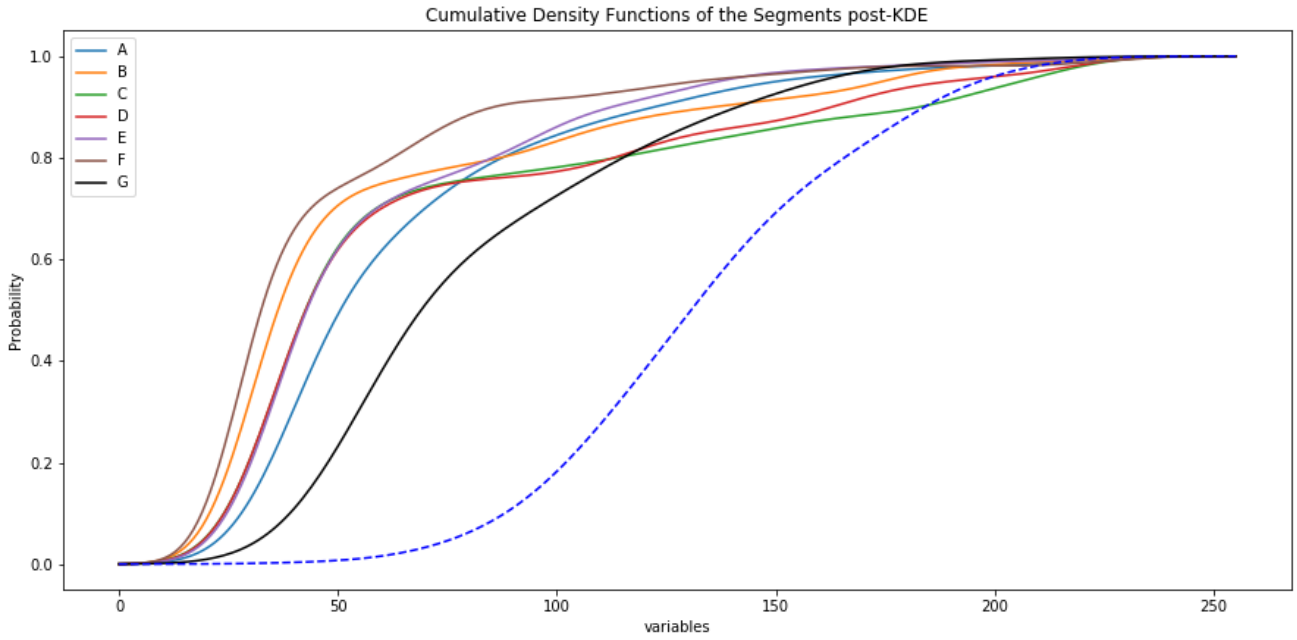


Figure 9: Cumulative distribution functions for all segments: each colour shows the specific segment, the blue dotted line depicts the Gaussian CDF.

For our analysis the use of the Archimedean family will better fit loan default dependencies as shown from Crook and Moreira (7) for credit risk modelling.
The copulas considered are: Gaussian, Clayton and Gumbel.

Step 2: Creation of Copulas

From section 3 we computed our variables defined as relative loss rates, with the application of KDE we obtained the uniform variables for all segments (see figure 9).
The density of each copula is derived from their density functions, for the Clayton copula:

$$C(u_1, u_2; \alpha) = (\alpha + 1)(u_1^{-\alpha} + u_2^{-\alpha} - 1)^{-2-(1/\alpha)} u_1^{-\alpha-1} u_2^{-\alpha-1}. \tag{11}$$

Where α is the parameter, u_1 and u_2 the uniform variables found from the KDE. For the Gaussian copula we used:

$$C(u_1, u_2; \varrho) = \Phi(\Phi^{-1}(u_1), \Phi^{-1}(u_2)). \tag{12}$$

Differently from the Clayton the Gaussian has ϱ as parameter and again u_1 and u_2 are the uniform variables estimated. Finally for the Gumbel copula we used the EQ(8) defined in section 4.
Each pair of segment (overall 21) was used to compute the copula densities according to the formulas (EQ(11) and EQ(12)). For the Clayton copula the application was straightforward since the density is in terms of the uniform however, for the Gaussian we applied first the inverse transformation

$(\phi^{-1}(u) = \varepsilon)$ to the uniform variables where the final density is computed by:

$$C(u_1, u_2; \varrho) = ln[((1-\varrho^2)^{-\frac{1}{2}}) * exp(-(2(1-\varrho^2))^{-1} * (\varrho^2\varepsilon_1^2\varepsilon_2^2) - 2(\varrho\varepsilon_1\varepsilon_2))]. \tag{13}$$

The following paragraph highlights the parametrisation procedure of α and ϱ for the two main copulas analysed.

Step 3: Parameter Estimation Technique

Parameter estimation can be obtained by three different common calibration methods. Maximizing a function that includes the parameters for both marginals and copula known as full maximum likelihood (3). Inference functions on margins (IFM) that maximises two likelihood functions first finding parameters of margins and then used to identify the copula parameters (7).

While full maximum likelihood (MLE) may be considered less transparent than IFM however, IFM could present less efficient estimators than MLE (3). Finally, the canonical maximum likelihood (CML) that is based on the estimation of parameters by maximising the log-likelihood function that includes uniform variables from the dataset and copula parameters (7).

After parametric estimation of the marginals by kernel density estimation (KDE) in section 3 ML is applied to estimate the parameters of the copula function. IFM has been criticised by the literature ((7) and (3)) because subject to flawed estimations during the "two-step" parametrisation. Our approach of estimating the margins through KDE and using ML from there tends to the CML method because it avoids the first parametrisation by applying the ML to the uniform variables from the dataset generated.

In practice to implement CML we need to work with the copula densities (15). The log-likelihood function can be explicitly defined as:

$$lnL(\alpha, \theta; x_1, ...x_T) = \sum_{t=1}^{T}(lnc(F_1(x_{1t}; \alpha_1), ..., F_n(x_n; \alpha_n); \theta) + \sum_{i=1}^{n} lnf_i(x_{it}; \alpha_i)) \tag{14}$$

Where marginal densities and distributions are $f_i(x_i; \alpha_i)$ and $F_i(x_i; \alpha_i)$, with $x_t = (x_{it}, ..., x_{nt})$ is the vector of observations on n random variables at time t in a sample of time series on the variables (3).

The log-likelihood can be expressed in terms of the copula densities for the case with dimensions n=2 by:

$$lnL(\theta; c_1, c_2) = ln \sum_{i=1}^{N}(c_1(u_1, ..., u_n), c_2(u_1, ..., u_n); \theta) \tag{15}$$

From full MLE the maximization involves more parameters than one-step ML, for higher dimensions it may become computationally cumbersome. CML permits to choose the best parameters by not being affected by the parametric estimation of the marginals (3).

Step 4: Model Selection Technique

After parametrization, the final step consist in selecting the best copula model. The major tech-

niques are: comparison of the results from the maximum likelihood function is the simplest one but subject to possible complications when parameters increase, another method is using information criteria as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) where the lowest value represents the best fit (3). All these methods share the invariance property under monotone increasing transformations of the marginals (16).

Finally, the measure that we consider in this analysis is based on the Goodness of Fit between the copula and the Empirical copula. Goodness of Fit (GoF) measures the difference in distribution between the copula found from the parametrisation steps and the empirical data. This method presents better performance according to Crook and Moreira (7) from using Empirical copula rather than Kendall's transform or Rosenblatt's transform.

Empirical copula distribution function can be defined as (3): Let N denote the number of pairs $(x, y)|x \leq x^i$ and $y \leq y^j$. Then set $C\left(\frac{i}{T}, \frac{j}{T}\right) = \frac{N}{T}$. Where $x_{(i)}$ and $y_{(j)}$ are order statistics from the sample.

To select the best copula based on GoF we compute the root mean square error (RMSE), defined by:

$$RMSE = \sqrt{\frac{1}{2}(c_1 - c_2)^2} \tag{16}$$

The model with the smallest RMSE gives the best copula for our data.

6 Empirical Results

The initial interpretation of the variables in section 3 gives us at a first glance the overall behaviour of the relative loss rates created where we see a slightly higher risk in the segment A than B (see table 1). All the consecutive segments presents an increasing mean according to the logic of risk reward trade-off. The table presents significant gaps between the RL rates mean over the period considered, pointing out that Lending Club is not effectively grading the credit segments since the gap widens between the segments such as F and G.

From section 3 we computed the relative loss rates (RL) per segment, the observed variables were then used for creating the uniform variables by kernel density estimation (see section 5). Working with the uniform variables facilitated two main steps of the copula application process: the first one consists into taking as input the uniform variables within the density function of each copula (e.g.: Calyton) in order to find the complete density of the RL rates dependence, the second for the selection of the best copula avoiding the "two-steps" maximisation issues discussed in section 5.

For the Clayton copula density function for two variables we used the equation EQ(11) and for the Gaussian copula we used the uniform variables transformed with EQ(12) in section 5.

The densities were then computed for each pair of the credit segment of LC thus, total of 21 pairs overall. The RL rates transformed into uniform variables were used to maximise the log-likelihood as seen in section 5. The diagonal of the matrix for the density was taken and the CML applied,[17]

[17] CML applied to the log-sum of the diagonal from the density function created for each variable: diagonal of the matrix 104x104

this process was carried out for all pairs in order to create the best fitted copula for our data by calibrating the parameters (α and ρ).
The following table shows the fitted copula parameters for Clayton copula α.

B	4.26					
C	3.66	9.84				
D	7.30	10.17	44.13			
E	10.29	8.63	14.19	18.89		
F	3.35	13.76	5.00	5.20	5.74	
G	2.84	1.39	1.68	1.79	1.98	1.23
	A	B	C	D	E	F

Table 2: α values per pair, the red cells represent the maximum and minimum of the array

The next table presents the fitted Gaussian copula parameters for ρ which ranges between 1 and 0 for perfect correlation and uncorrelated variables respectively.

B	0.959					
C	0.928	0.953				
D	0.954	0.973	0.999			
E	0.985	0.990	0.898	0.929		
F	0.924	0.960	0.877	0.902	0.959	
G	0.926	0.843	0.8333	0.862	0.886	0.778
	A	B	C	D	E	F

Table 3: ρ values per pair, the blue cells represent the maximum and minimum of the array

As expected we see significantly high levels of correlations between the pairs of segments, all higher than 77%. The only pair with lower correlation coefficient is F-G, this could lead to a misinterpretation of the risk level for the two segments where linear correlation might not be appropriate. Indeed, the best copula selected is the Clayton with the lowest parameter for α. On the other hand the extreme value of correlation for the pair D-C around 0.999 shows the possibility of using the correct measure of risk since the RL rates follow each other with the Gaussian copula however, the GoF selection gives us Clayton copula as better option. The assumption of avoiding the Gaussian should be considered but always used as benchmark for comparison because it can lead to effective risk interpretations: for the pair F-G the difference between the Gaussian and Clayton is significantly small (referring to the RMSE of all pairs).[18] Overall the Clayton copula is dominant in all cases and the result of using it will facilitate risk evaluations and predictions.
Once the fitted copula was computed and the parameters calibrated we proceeded to the copula selection with the method of Goodness-of-Fit (see section 5). The Empirical Copula was created by dividing equally the sum of the number of variables (104) over the whole density of the function, figure 10 depicts the case of the segments A-B Empirical Copula density.

[18] see section 5 for formula

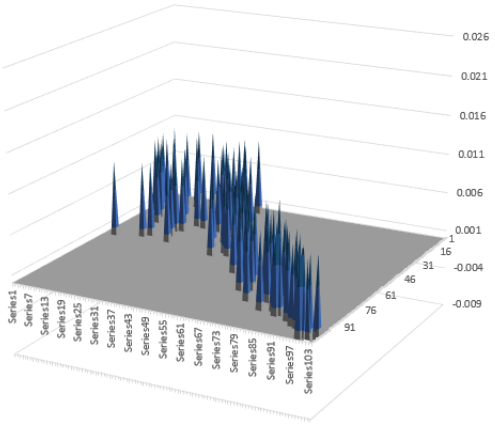
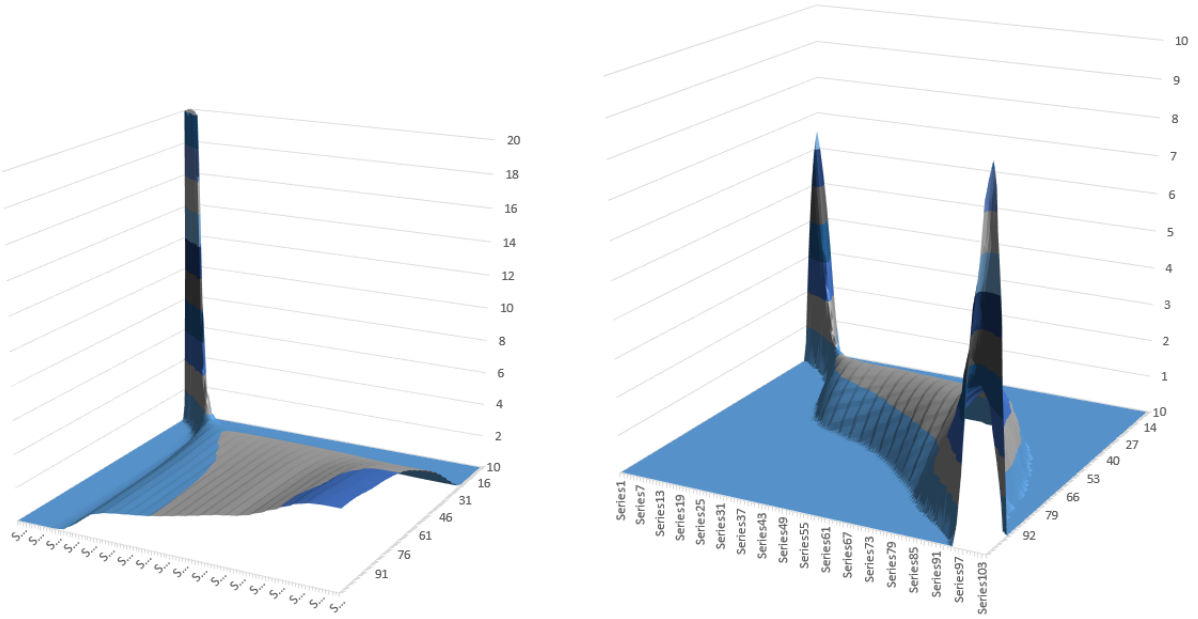


Figure 10: Example of Empirical Copula for the segment pair: A-B

After the parametrisation by ML the copula selection is carried out by comparison between the two main copula functions analysed: Gaussian and Clayton. To choose the best copula we compared the RMSE values according to the GoF procedure (see section 5) where the smallest value highlights the copula function that better capture the extreme events in the tails.
After the GoF selection we see how for all pairs the Clayton copula captures better the extreme tails, since the whole distribution is considered the Gaussian can be seen as competitive model however, further studies may focus more on the tail dependence and see how Gaussian is not appropriate according to the findings in the literature review section (2).
As expected the majority of copula functions that properly fit the data is the Clayton, the extreme tail movements are captured by this copula that considers the jumps in relative loss rates better than Gaussian copula model for lower tail dependencies.

Figure 11 part (a) represents the density of the Clayton copula fitted for our pair (A-C in this scenario), part (b) shows on the other hand the density of the Gaussian copula.



(a) Example of Clayton copula density for the segment pair: A-C; $\alpha = 3.66$

(b) Example of Gaussian copula density for the segment pair: A-C; $\rho = 0.928$

Figure 11: Densities of Clayton and Gaussian copulas

We can see how the extreme events of the Empirical Copula will be better captured by the Clayton copula that presents lower tail dependence.
Finally for complete comparison we plot the Gumbel fitted copula in figure 12 for the same pair. It captures as opposite of the Clayton copula the upper tail default dependence. For this reason we decided to not include the copula in further analysis since the empirical copulas present higher density within the lower tail.

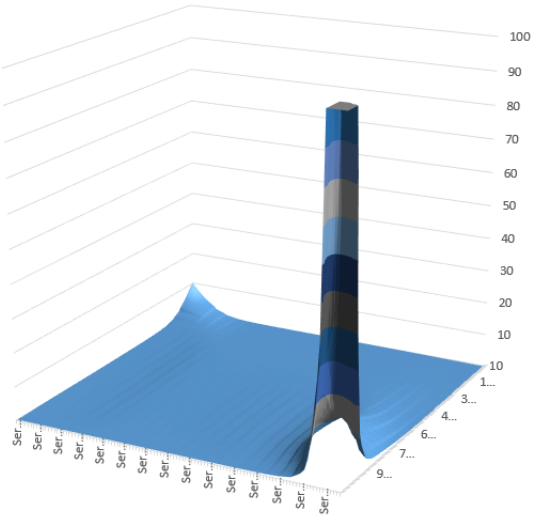


Figure 12: Example of Gumbel copula density for the segment pair: A-C; $\delta = 1.365$

Overall all pairs behave similarly for exception of the combination F-G and C-D. The first one has particular characteristics (see section 3 for volume issue) where instead the latter (C-D) seems almost perfectly correlated and tending to the Gaussian copula more than all the other pairs nonetheless, Clayton remains the best choice.

7 Conclusion

This report analyses the Lending Club data and gives an overview of credit risk within P2P lending platforms. It shows how risk is managed within the FinTech sector by stressing the importance of appropriate credit scoring and transparency to the investors.

From section 3 we see how our assumptions on the sensitivity to macroeconomic dynamics for loan defaults affect relative loss rates for LC less than default rates. Moreover, the use of copulas for default dependencies shows the effectiveness of building a correct model by following the steps listed in section 5 where credit risk is captured by the Clayton copula. Finally, the comparison and selection sections confirm the intuition from section 1 of the inadequate usage of Gaussian copulas for credit risk measurement.

In the section 3 the construction of relative loss rates in comparison to default rates provides less sensitivity to macroeconomic changes. However, RL rates are less effective with a lower number of loans because small changes in the volume will heavily affect the rates creating biased values. The period 2007-2009 was showing RL rates close to 50% for low risk profile segments (A and B) thus, completely unbalanced and defined outliers for our analysis.

The analysis is taking the whole distribution of the dataset in consideration however, a possible improvement would lead to a specific investigation of tail dependencies. In particular, for the C-D segment pair case (see section 6), possible different results with the application of only the tails could bring the acceptance of the Gaussian instead of Clayton copula.

Finally, further studies should carry out a more extensive analysis by considering P2P lending platforms within macroeconomic financial crisis events during their operations even with a wider timespan. Our condition of LC as established FinTech business helps in determining a solid reference rate for future analysis where different extreme events could test the models used here for measuring credit risk.

References

[1] Altman, E. I. (November 1980). *Commercial Bank Lending: Process, Credit Scoring, and Costs of Errors in Lending* The Journal of Financial and Quantitative Analysis, Vol. 15, No. 4, Proceedings of California, pp. 813-832.

[2] Alexander, C. (2008) *Quantitative Methods In Finance* John Wiley & Sons, Vol. 1, Section I.3.3.12.

[3] Alexander, C. (2008) *Practical Financial Econometrics* John Wiley & Sons, Vol. 2, Section II.6.

[4] Cherubini, U. & Luciano, E. (2002). *Bivariate option pricing with copulas* Applied Mathematical Finance, 9, 69-85.

[5] Cherubini, U. et alt. (2012) *Dynamic Copula Methods in Finance* John Wiley & Sons, Section II.

[6] Clayton, D. (1978) *A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence.* Biometrika 65, 141–151.

[7] Crook, J. & Moreira, F. (2011). *Checking for asymmetric default dependence in a credit card portfolio: A copula approach* Journal of Empirical Finance, 18, 728-742.

[8] Deloitte UK. *A temporary phenomenon? Marketplace lending an analysis of the UK market https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-fs-marketplace-lending.pdf*

[9] Dematra, S. & McNeil, A. J. (2005). *The t Copula and Related Copulas* International Statistical Review, 73, 1, 111-129 (2005).

[10] Fenech, J. P. et alt. (2015) *Loan default correlation using an Archimedean copula approach: A case for recalibration* Economic Modelling, 47, 340-354.

[11] Jarque, C.M. & Bera, A.K. (1987) *A Test for Normality of Observations and Regression Residuals* Internal Statistical Review, 55, 163-172.

[12] Joseph, C. (2013) *Advanced Credit Risk Analysis and Management* John Wiley & Sons, Section III.

[13] Li, D. X. (March 2000). *On Default Correlation: A Copula Function Approach* The Journal of Fixed Income, Business Premium Collection, pg. 43.

[14] MacKenzie, D. & Spears, T. (June 2012). *"The Formula That Killed Wall Street?" The Gaussian Copula and the Material Cultures of Modelling.* University of Edinburgh.

[15] McNeil, A. J. et alt. (2005) *Quantitative Risk Management: Concepts, techniques and Tools* Princeton University Press, Section V.

[16] Silva, R. d. S. & Lopes, H. F. (2008). *Copula, marginal distributions and model selection: a Bayesian note*. 18: 313-320, Stat Comput. *https://doi.org/10.1007/s11222-008-9058-y*

[17] Silverman, B. (November 1987) *Density Estimation for Statistics and Data Analysis*. London, Chapman & Hall, Vol. 29, No. 4, p. 495.

[18] Sklar, A. (1996). *Random Variables, Distribution Functions, and Copulas: A Personal Look Backward and Forward* Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Vol. 28, Distributions with Fixed Marginals and Related Topics, pp. 1-14 .

[19] Zhang, B. et alt. (December 2017). *Entrenching Innovation: The 4th UK Alternative Finance Industry Report*. Cambridge Centre for Alternative Finance, University of Cambridge Judge Business School.

[20] Zheng, Y. et alt. (2011) *Approximation of bivariate copulas by patched bivariate Frechet copulas* Insurance: Mathematics and Economics, 48, 246-256.