Factorial Network Models To Improve P2P Credit Risk Management

Boston University - Department of Mathematics and Statistics, University of Pavia - Faculty of Economics, ZHAW University of Applied Sciences

26 February 2019

Factorial Network Models To Improve P2P Credit Risk Management

Daniel Felix Ahelegbey[a,*], Paolo Giudici[1], Branka Hadji-Misheva[c]

[a] *Department of Mathematics and Statistics, Boston University, USA*
[b] *Department of Economics and Management, University of Pavia, Italy*
[c] *ZHAW University of Applied Sciences, Zurich, Switzerland*

Abstract

This paper investigates how to improve statistical-based credit scoring of SMEs involved in P2P lending. The methodology discussed in the paper is a factor network-based segmentation for credit score modeling. The approach first constructs a network of SMEs where links emerge from comovement of latent factors, which allows us to segment the heterogeneous population into clusters. We then build a credit score model for each cluster via lasso-type regularization logistic regression. We compare our approach with the conventional logistic model by analyzing the credit score of over 15000 SMEs engaged in P2P lending services across Europe. The result reveals that credit risk modeling using our network-based segmentation achieves higher predictive performance than the conventional model.

Keywords: Credit Risk, Factor models, Fintech, Peer-to-Peer lending, Credit Scoring, Lasso, Segmentation

1. Introduction

Issuance of loans by traditional financial institutions, such as banks, to other firms and individuals, is often associated with major risks. The failure of loan recipients to honor their obligation at the time of maturity leaves the banks vulnerable and affects their operations. The risk associated with such transactions is referred to as credit risk. It is well known that some percentage of these non-performing loans are eventually imputed to economic losses. To minimize such risk exposures, various methods have been extensively discussed in the credit risk literature to enable credit-issuing institutions to undertake a thorough assessment to classify loan applicants into risky and non-risky customers. Some of these methods range from logistic and linear probability models to decision trees, neural networks and support vector machines. A conventional individual-level reduced-form approach is the credit scoring model which attributes a score of credit-worthiness to each loan applicant based on the available history of their financial characteristics. See Altman (1968) for some pioneer works on corporate bankruptcy prediction models using accounting-based measures as variables. For a comprehensive review on credit scoring models, see Alam et al. (2010).

Recent advancements gradually transforming the traditional economic and financial system is the emergence of digital-based systems. Such systems present a paradigm shift from

*Corresponding author at: Department of Mathematics and Statistics, Boston University, USA.
Email addresses: dfkahey@bu.edu (Daniel Felix Ahelegbey), paolo.giudici@unipv.it (Paolo Giudici), branka.hadjimisheva01@universitadipavia.it (Branka Hadji-Misheva)

traditional infrastructural systems to technological (digital) systems. Financial technological ("FinTech") companies are gradually gaining ground in major developed economies across the world. The emergence of Peer-to-Peer (P2P) platforms is a typical example of a FinTech system. The P2P platform aims at facilitating credit services by connecting individual lenders with individual borrowers without the interference of traditional banks as intermediaries. Such platform serves as a digital financial market and an alternative to the traditional physical financial market. P2P platforms significantly improve the customer experience and the speed of the service and reduce costs to both individual borrowers and lenders as well as small business owners. Despite the various advantages, P2P systems inherit some of the challenges of traditional credit risk management. In addition, they are characterized by the asymmetry of information and by a strong interconnectedness among their users (see e.g. Giudici et al., 2019) that makes distinguishing healthy and risky credit applicants difficult, thus affecting credit issuers. There is, therefore, a need to explore methods that can help improve credit scoring of individual or companies that engage in P2P credit services.

This paper investigates how factor-network-based segmentation can be employed to improve the statistical-based credit score for small and medium enterprises (SMEs) involved in P2P lending. The approach is to first constructs a network of SMEs where links emerge from comovement of the latent factors that drive the observed financial characteristics. The network structure then allows us to segment the heterogeneous population into two subgroups of connected and non-connected clusters. We then build a credit score model for each sub-population via lasso-type regularization logistic regression.

The contribution to the literature of this paper is manifold. Firstly, we extend the ideas contained in the factor network-based classification of Ahelegbey et al. (2019) to a more realistic setting, characterized by a large number of observations which, when links between them are the main object of analysis, becomes extremely challenging.

Secondly, we extend the network-based scoring model proposed in Giudici et al. (2019) to a setting characterized by a large number of explanatory variables. The variables are selected via lasso-type regularization (Tibshirani, 1996; Trevor et al., 2009) and, then, summarized by factor scores. Thus, we contribute to network-based models for credit risk quantification. Network models have been shown to be effective in gauging the vulnerabilities among financial institutions for risk transmission (see Ahelegbey et al., 2016a; Battiston et al., 2012; Billio et al., 2012; Diebold and Yilmaz, 2014), and a scheme to complement micro-prudential supervision with macro-prudential surveillance to ensure financial stability (see IMF, 2011; Moghadam and Viñals, 2010; Viñals et al., 2012). Recent application of networks have been shown to improve loan default predictions and capturing information that reflects underlying common features (see Ahelegbey et al., 2019; Letizia and Lillo, 2018).

Thirdly, our empirical application contributes to modeling credit risk in SMEs particularly engaged in P2P lending. For related works on P2P lending via logistic regression, see Andreeva et al. (2007); Barrios et al. (2014); Emekter et al. (2015); Serrano-Cinca and Gutiérrez-Nieto (2016). We model the credit score of over 15000 SMEs engaged in P2P credit services across Southern Europe. We compare the performance of our network-based segmentation credit score model (NS-CSM) with the conventional single credit score model (CSM). We show via our empirical results that our network-based segmentation presents a more efficient scheme that achieves higher performance than the conventional approach.

The paper is organized as follows. Section 2 presents the factor network segmentation methodology and the lasso-type regularization for credit scoring. Section 3 discusses the empirical application of our segmentation approach against the conventional single model.

2. Methodology

We present the formulation and inference of a latent factor network to improve credit scoring and model estimation. Our objective is to analyze the characteristics of the borrowers to build a model that predicts the likelihood of their default.

2.1. Logistic Model

Let Y be a vector of independent observations of the loan status of n firms, such that $Y_i = 1$ if firm-i has defaulted on its loan obligation, and zero otherwise. Furthermore, let $X = \{X_{ij}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, be a matrix of n observations with p financial characteristic variables or predictors. The conventional parameterization of the conditional distribution of Y given X is the logistic model with log-odds ratio given by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + X_i\beta \tag{1}$$

where $\pi_i = P(Y_i = 1|X_i)$, β_0 is a constant term, $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of coefficients and X_i is the i-th row of X.

2.2. Decomposition of Data Matrix by Factors

The dataset X can be considered as points of n-institutions in a p-dimensional space. It can also be interpreted at observed outcomes driven by some underlying firm characteristics. More specifically, X can be expressed as a factor model given by

$$X \;=\; FW + \varepsilon \tag{2}$$

where F is $n \times k$ matrix of latent factors, W is $p \times k$ matrix of factor loadings, ε is $n \times p$ matrix of errors uncorrelated with F. The error term ε is typically assumed to be multivariate normal but F in general case need not be multivariate normal (see Tabachnick et al., 2007). Lastly, $k < p$ is the number of factors required to summarize the pattern of correlations in the observed data matrix X. In the context of our application, we set k to be the number of factors that account for approximately 95% of the variation in X.

2.3. Factor Network-Based Segmentation

We present the construction of network structure for the segmentation of the population. Following the literature on graphical models (see Ahelegbey et al., 2016a,b; Carvalho and West, 2007; Eichler, 2007), we represent the network structure as an undirected binary matrix, $G \in \{0,1\}^{n\times n}$, where G_{ij} represents the presence or absence of a link between nodes i and j. We construct G via similarity of the latent firm characteristics, such that $G_{ij} = 1$ if the latent coordinates of firm-i are strongly related to firm-j, and zero otherwise.

Given the latent factors matrix, F, we construct a network where the marginal probability of a link between nodes-i and j by

$$\gamma_{ij} \;=\; P(G_{ij} = 1|F) \;=\; \Phi[\theta + (FF')_{ij}] \tag{3}$$

where $\gamma_{ij} \in (0,1)$, Φ is the standard normal cumulative density function, $\theta \in \mathbb{R}$ is a network density parameter, and $(FF')_{ij}$ is the i-th row and the j-th column of FF'. Under the

assumption that G is undirected, it follows that $\gamma_{ij} = P(G_{ij} = 1|F) = P(G_{ji} = 1|F) = \gamma_{ji}$. We validate the link between nodes-i and j in G by

$$G_{ij} \;=\; \mathbf{1}(\gamma_{ij} > \gamma) \tag{4}$$

where $\mathbf{1}(\gamma_{ij} > \gamma)$ is the indicator function, i.e., unity if $\gamma_{ij} > \gamma$ and zero otherwise, and $\gamma \in (0, 1)$ is a threshold parameter. By definition, the parameters θ and γ control the density of G. Following Ahelegbey et al. (2019), we set $\theta = \Phi^{-1}(\frac{2}{n-1})$. To broaden the robustness of the results, we compare $\gamma = \{0.05, 0.1\}$ to capture a sparse but closely connected community.

2.4. Estimating High-Dimensional Logistic Models

When estimating high-dimensional logistic models with a relatively large number of predictors, there is the tendency to have redundant explanatory variables. Thus, to construct a predictable model, there is the need to select the subset of predictors that explains a large variation in the probability of defaults. Several variable selection methods have been discussed and applied for various regression models. In this paper, we consider variants of the lasso regularization for logistic regressions (Trevor et al., 2009).

2.4.1. Lasso

The lasso estimator (Tibshirani, 1996) solves a penalized log-likelihood function given by

$$\arg\min_{\beta} \sum_{i=1}^{n} \Big[Y_i(\beta_0 + X_i\beta) - \log\left(1 + \exp(\beta_0 + X_i\beta)\right) \Big] - \lambda \sum_{j=0}^{p} |\beta_j| \tag{5}$$

where n is the number of observations, p the number of predictors, and λ is the penalty term, such that large values of λ shrinks a large number of the coefficients towards zero.

2.4.2. Adaptive Lasso

The adaptive lasso estimator (Zou, 2006) is an extension of the lasso that solves

$$\arg\min_{\beta} \sum_{i=1}^{n} \Big[Y_i(\beta_0 + X_i\beta) - \log\left(1 + \exp(\beta_0 + X_i\beta)\right) \Big] - \lambda \sum_{j=0}^{p} w_j|\beta_j| \tag{6}$$

where w_j is a weight penalty such that $w_j = 1/|\hat{\beta}_j|^v$, with $\hat{\beta}_j$ as the ordinary least squares (or ridge regression) estimate and $v > 0$.

2.4.3. Elastic-Net

The elastic-net estimator (Zou and Hastie, 2005) solves the following

$$\arg\min_{\beta} \sum_{i=1}^{n} \Big[Y_i(\beta_0 + X_i\beta) - \log\left(1 + \exp(\beta_0 + X_i\beta)\right) \Big] - \lambda \sum_{j=0}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2) \tag{7}$$

where $\alpha \in (0, 1)$ is an additional penalty such that when $\alpha = 1$ we a lasso estimator (L_1 penalty), and when $\alpha = 0$ a ridge estimator (L_2 penalty). For the elastic-net estimator, we set $\alpha = 0.5$ giving equal weight to the L_1 and L_2 regularization.

2.4.4. Adaptive Elastic-Net

The adaptive elastic-net estimator (Zou and Zhang, 2009) combines the additional penalties of the adaptive lasso and the elastic-net to solve the following

$$\arg\min_{\beta} \sum_{i=1}^{n} \Big[Y_i(\beta_0 + X_i\beta) - \log\left(1 + \exp(\beta_0 + X_i\beta)\right) \Big] - \lambda \sum_{j=0}^{p} (\alpha w_j |\beta_j| + (1-\alpha)\beta_j^2) \tag{8}$$

In the empirical work, we focus on estimating the credit score using the four lasso-type regularization methods. We select the regularization parameter using ten-fold cross-validation on a grid of λ values for the penalized logistic regression problem. Two λ's are widely considered in the literature, i.e., $\lambda.min$ and $\lambda.1se$. The former is the value of the λ that minimizes the mean square cross-validated errors, while the latter is the λ value that corresponds to one standard error from the minimum mean square cross-validated errors. Our preliminary analysis shows that $\lambda.1se$ produces a larger penalty that is too restrictive in the sense that we lose almost all the regressors. Although our goal is to encourage a sparse credit scoring model for the purpose of interpretability, we do not want to impose too much sparsity that renders the majority of the features insignificant. Thus, we rather choose $\lambda.min$ over $\lambda.1se$. For the additional penalty terms, we set $\alpha = 0.5$, $v = 2$, and $\hat{\beta}_j$ as the ridge regression estimate.

3. Application

3.1. Data: Description and Summary Statistics

To illustrate the effectiveness of the application of factor network methodology in credit scoring analysis, we obtained data from the European External Credit Assessment Institution (ECAI) on 15045 small-medium enterprises engaged in Peer-to-Peer lending on digital platforms across Southern Europe. The observation on each institution is composed of 24 financial characteristic ratios constructed from official financial information recorded in 2015. Table 1 presents a description of the financial ratios with summary of mean statistics of the institutions grouped according to their default status. In all, the data consists of 1,632 (10.85%) defaulted institutions and 13,413 (89.15%) non-defaulted companies.

3.2. Decomposition of the Observed Data Matrix by Factors

To estimate the underlying factors that drive the observed data matrix, we decompose the matrix of observed financial characteristics via a singular value decomposition given by,

$$X \;=\; UDV \;=\; FW + \varepsilon \tag{9}$$

where U and V are orthonormal, and $D = \Lambda^{1/2}$ is a diagonal matrix of non-negative and decreasing singular values, with Λ as the diagonal matrix of the non-zero eigenvalues of $X'X$ and XX'. U is $n \times p$, D is $p \times p$ and V is $p \times p$. Following the error approximation criteria, we obtain the factor matrix by, $F = U_{n,k}\ D_{k,k}$ and $W = V_{k,p}$, where $U_{n,k}$ is $n \times k$ matrix composed of the first k columns of U, $k < p$, $D_{k,k}$ is $k \times k$ matrix comprising the first k columns and rows of D, and $V_{k,p}$ is $k \times p$ matrix of factor loadings. The matrix F can therefore be interpreted as a projection of X onto the eigenspace spanned by $U_{n,k}$. We determine k by observing the number of eigenvalues associated with the largest variance matrix. Table 2 shows the eigenvalues of the singular value decomposition to determine the factors to retain. The eigenvalues reported are the normalized squared diagonal terms of D. From the table, we set $k = 17$ since the first 17 eigenvalues explain about 95% of the total variation in X.

Var	Formula (Description)	Active(Mean)	Defaulted(Mean)
V1	(Total Assets - Shareholders Funds)/Shareholders Funds	8.87	9.08
V2	(Longterm debt + Loans)/Shareholders Funds	1.25	1.32
V3	Total Assets/Total Liabilities	1.51	1.07
V4	Current Assets/Current Liabilities	1.6	1.06
V5	(Current Assets - Current assets: stocks)/Current Liabilities	1.24	0.79
V6	(Shareholders Funds + Non current liabilities)/Fixed Assets	8.07	5.99
V7	EBIT/Interest paid	26.39	-2.75
V8	(Profit (loss) before tax + Interest paid)/Total Assets	0.05	-0.13
V9	P/L after tax/Shareholders Funds	0.02	-0.73
V10	Operating Revenues/Total Assets	1.38	1.27
V11	Sales/Total Assets	1.34	1.25
V12	Interest Paid/(Profit before taxes + Interest Paid)	0.21	0.08
V13	EBITDA/Interest Paid	40.91	5.71
V14	EBITDA/Operating Revenues	0.08	-0.12
V15	EBITDA/Sales	0.09	-0.12
V16	Constraint EBIT	0.13	0.56
V17	Constraint PL before tax	0.16	0.61
V18	Constraint Financial PL	0.93	0.98
V19	Constraint P/L for period	0.19	0.64
V20	Trade Payables/Operating Revenues	100.3	139.30
V21	Trade Receivables/Operating Revenues	67.59	147.12
V22	Inventories/Operating Revenues	90.99	134.93
V23	Total Revenue	3557	2083
V24	Industry Classification on NACE code	4566	4624
	Total number of institutions (%)	13413 (89.15%)	1632 (10.85%)

Table 1: Description of the financial ratios with summary of mean statistics according to default status.

3.3. Factor Network Analysis

We use the estimated factor matrix, F, to construct the network for the segmentation of the companies. For purposes of graphical representations and to keep the companies name anonymous, we report the estimated network by representing the group of institutions with color-codes. The defaulted companies are represented in a red color code, and non-defaulted companies in the green color code (see Figure 1). Table 3 reports the summary statistics of the estimated network in terms of the default-status composition of the SMEs. For robustness purposes, we compare the results obtained with a threshold value $\gamma = 0.05$ against $\gamma = 0.10$.

The result for the threshold $\gamma = 0.05$ of Table 3 shows that the connected sub-population is composed of 4305 companies which constitute 28.6% of the full sample. The non-connected sub-population is composed of 10740 (71.4%). The percentage of the defaulted class of companies are 22.4% and 6.2% among the connected- and non-connected sub-population, respectively. We notice that higher threshold values (say $\gamma = 0.1$) decrease (increase) the total number of connected (non-connected) sub-population and vice versa. Such higher threshold values also lead to a lower (higher) number of defaulted class of connected (non-connected) SMEs but (and) constituting a higher percentage of the defaulted population. Figure 1 presents the graphical representation of the estimated factor network with the sub-population of defaulted and non-defaulted companies color coded as red and green, respectively. Figure 1a shows the structural representation of both connected and non-connected sub-population while Figure 1b depicts the structure of connected sub-population only.

No.	Eigenvalue	Variance Explained (%)	Cumulative (%)
1	5.18	21.60	21.60
2	2.58	10.73	32.33
3	2.50	10.41	42.74
4	1.60	6.69	49.42
5	1.42	5.92	55.34
6	1.30	5.40	60.74
7	1.16	4.82	65.55
8	1.09	4.56	70.11
9	0.99	4.11	74.22
10	0.93	3.88	78.10
11	0.80	3.35	81.45
12	0.79	3.31	84.76
13	0.75	3.11	87.87
14	0.56	2.35	90.22
15	0.53	2.21	92.43
16	0.51	2.12	94.55
17	0.43	1.80	96.35
18	0.37	1.54	97.89
19	0.17	0.69	98.58
20	0.11	0.47	99.05
21	0.09	0.36	99.41
22	0.07	0.27	99.68
23	0.06	0.26	99.94
24	0.01	0.06	100.00

Table 2: The eigenvalues of the singular value decomposition to determine the factors to retain.

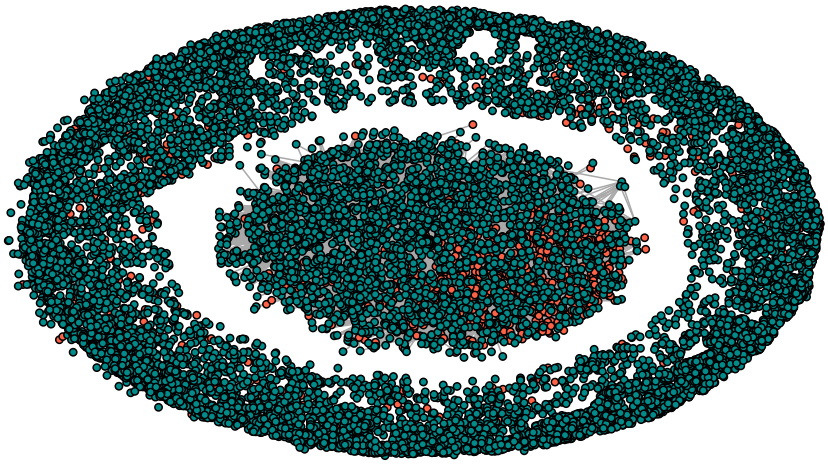
Threshold	Status	Conn-Sub	Non-Conn-Sub
$\gamma = 0.05$	Default	964 - 22.4%	668 - 6.2%
	Non-Default	3,341 - 77.6%	10,072 - 93.8%
	Total	4,305 - 28.6%	10,740 - 71.4%
$\gamma = 0.1$	Default	816 - 24%	816 - 7%
	Non-Default	2,580 - 76%	10,833 - 93%
	Total	3,396 - 22.6%	11,649 - 77.6%

Table 3: Summary statistic of connected and non-connected sub-population obtained from the factor network-based segmentation for threshold values of $\gamma = \{0.05, 0.1\}$.

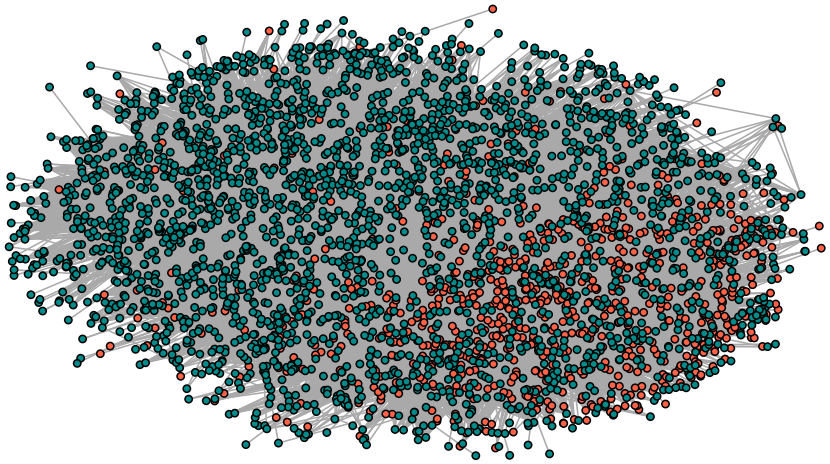
3.4. Credit Score Modeling

We compare the lasso, adaptive lasso, elastic-net, and adaptive elastic-net variable selection methods to model the credit score of the listed companies in our dataset. To estimate the models, we standardized each series to a zero mean and unit variance. Table 4 reports the variable selection and estimated coefficients of the four methods. The column CSM represents the benchmark credit scoring model, NS-CSM(C) - the network segmented connected sub-population credit scoring model, and NS-CSM(NC) for the network segmented non-connected sub-population credit scoring model. The top left panel represents the lasso method, the adaptive lasso is on the top right panel, elastic-net at the bottom left and adaptive elastic-net at the bottom right.

Table 5 reports the number of variables selected by each of the four competing methods for the credit score model estimation. From the table, the elastic-net is the least parsimonious,



(a) Network Structure of All Institutions



(b) Network of Connected Component

Figure 1: A graphical representation of the estimated factor network. (1a) shows the structural representation of the factor network for threshold $\gamma = 0.05$, and (1b) depicts the connected sub-population only. The nodes in red-color are defaulted class of companies and green-color coded nodes are non-defaulted class of companies.

followed by the lasso, and lastly, the adaptive elastic-net and adaptive lasso are the most parsimonious. From Tables 4 and 5, we observed a significant difference in the number of selected explanatory variables for the benchmark model and the network segmented models. More precisely, the former model the credit score of a given company by using more variables while the latter on the other hand uses a significantly lower number of variables. The similar results across the four variable selection methods, given their similarities, is not terribly surprising. But they do indicate that the general approach appears to be robust in this setting, which was the main purpose of the testing. The network-based segmentation framework is therefore more parsimonious than the benchmark full population credit score model, and this helps in interpretability.

3.5. Comparing Default Predicting Accuracy

We analyzed the performance of the models by splitting the sample into 70% training and 30% testing sample. We now compare the default prediction accuracy of the models in terms

	CSM	NS-CSM(C)	NS-CSM(NC)	CSM	NS-CSM(C)	NS-CSM(NC)
	Lasso			*Adaptive Lasso*		
V1	0.0535	.	0.0375	.	.	.
V2	.	0.0332	.	.	.	.
V3	-0.4468	-0.2818	-1.0148	-0.5298	-0.3539	-1.1990
V4	-0.3549	-0.1294	-0.5556	-0.2928	-0.1368	-0.5137
V5	.	.	.	.	.	.
V6	0.0774	.	0.1460	0.0440	.	0.0213
V7	0.2818	.	.	0.2116	.	.
V8	-0.3933	-0.3408	0.1185	-0.4356	-0.3463	.
V9	-0.0360	0.0365	-0.4690	.	.	-0.5577
V10	-0.0701	0.0287	.	.	.	.
V11	0.1291	.	0.0550	.	.	.
V12	0.0265	0.0222	0.0204	.	.	.
V13	-0.2419	.	.	-0.1759	.	.
V14	-0.0399	-0.0776	.	.	-0.113	.
V15	-0.0751	-0.0396	0.0128	-0.0520	.	.
V16	0.0520	0.2851	.	.	0.2245	.
V17	0.2213	0.1650	0.1761	0.2529	0.2092	.
V18	0.0396	0.0661	0.0143	.	0.0484	.
V19	0.2540	0.0291	0.2096	0.2755	.	0.2151
V20	0.0412	.	0.2429	.	.	0.1950
V21	0.2212	0.1620	0.2969	0.2410	0.1721	0.3185
V22	0.0930	.	0.1470	0.0541	.	0.0219
V23	-0.2262	-0.0649	-0.3452	-0.2213	-0.0650	-0.3826
V24	-0.0062	-0.0641	0.0343	.	-0.0645	.
	Elastic-Net			*Adaptive Elastic-Net*		
V1	0.0548	.	0.0568	.	.	.
V2	1.0e-04	0.0372	.	.	.	.
V3	-0.4472	-0.2692	-1.0132	-0.5293	-0.3538	-1.2208
V4	-0.3628	-0.1286	-0.6051	-0.2900	-0.1350	-0.6034
V5	0.0048	-0.0123	.	.	.	.
V6	0.0780	-0.0028	0.1862	0.0422	.	0.1528
V7	0.3003	.	.	0.1925	.	.
V8	-0.3926	-0.3310	0.2054	-0.4363	-0.3474	0.1672
V9	-0.0356	0.0435	-0.4884	.	.	-0.5195
V10	-0.1419	0.0315	.	.	.	.
V11	0.2016	0.0112	0.1025	.	.	.
V12	0.0299	0.0299	0.0545	.	.	.
V13	-0.2595	.	.	-0.1571	.	.
V14	-0.0374	-0.0785	.	.	-0.1112	.
V15	-0.0777	-0.0468	0.0597	-0.0499	.	.
V16	0.0600	0.2902	0.0669	.	0.2256	.
V17	0.2173	0.1588	0.1701	0.2527	0.2097	0.1147
V18	0.0417	0.0769	0.0439	.	0.0459	.
V19	0.2538	0.0502	0.2042	0.2747	.	0.2151
V20	0.0425	.	0.3139	.	.	0.2571
V21	0.2210	0.1634	0.3113	0.2409	0.1721	0.3036
V22	0.0933	0.0012	0.1727	0.0533	.	0.1047
V23	-0.2286	-0.0728	-0.3754	-0.2185	-0.0616	-0.4114
V24	-0.0077	-0.0724	0.0464	.	-0.0619	.

Table 4: Estimated coefficients from lasso (top left), adaptive lasso (top right), elastic-net (bottom left) and adaptive elastic-net (bottom right). CSM is the benchmark credit score model, NS-CSM(C) is the network segmented connected sub-population credit score model, and NS-CSM(NC) is the network segmented non-connected sub-population credit score model, estimated for threshold value $\gamma = 0.1$.

of the standard area under the curve (AUC) derived from the receiver operator characteristic (ROC) curve. The AUC depicts the true positive rate (TPR) against the false positive rate (FPR) depending on some threshold. TPR is the number of correct positive predictions divided by the total number of positives. FPR is the ratio of false positives predictions

	Lasso	Adaptive Lasso	Elastic-Net	Adaptive Elastic-Net
CSM	22	12	24	12
NS-CSM(C)	16	10	20	10
NS-CSM(NC)	17	9	18	11

Table 5: Number of selected variables of the four methods.

overall negatives. See Figure 2 for the plot of the ROC curve for the competing methods.

	Lasso	Adaptive Lasso	Elastic-Net	Adaptive Elastic-Net
CSM	0.8089	0.8061	0.8090	0.8061
NS-CSM($\gamma = 0.05$)	0.8214	0.8204	0.8225	0.8207
NS-CSM($\gamma = 0.1$)	0.8330	0.8277	0.8342	0.8312

Table 6: Comparing area under the ROC curve (AUC) of the four methods.

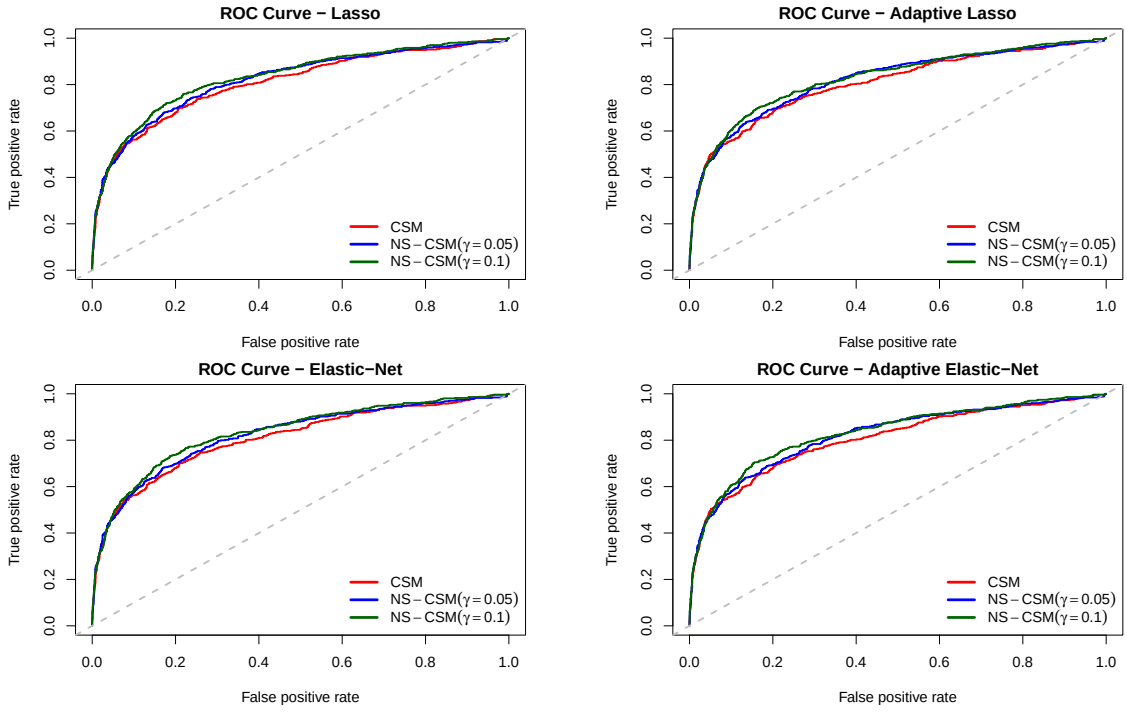


Figure 2: ROC curves of the four methods. CSM is the benchmark model, NS-CSM(C) is the network segmented connected sub-population model, and NS-CSM(NC) is the network segmented non-connected sub-population model, estimated for threshold values of $\gamma = \{0.05, 0.1\}$.

The comparison of the ROC curves from the competing methods shows that the CSM (in red) lies below the rest. Clearly, the curves of NS-CSM ($\gamma = 0.1$) depicted in green seems to dominate the others. The summary of the area under the ROC curve reported in Table 6 shows that NS-CSM ($\gamma = 0.1$) is ranked first, followed by NS-CSM ($\gamma = 0.05$), and the lowest AUC is obtained by the CSM. Overall, in terms of default predictive accuracy, the result of the AUC shows the NS-CSM outperforms the CSM, on average by two percentage points. This is an advantage that can be further increased considering as the cut-off the observed

default percentages, which are different in the two samples.

		Statistic	P-value	Significance	Statistic	P-value	Significance
			Lasso			*Adaptive Lasso*	
CSM	NS-CSM($\gamma = 0.05$)	-0.7639	0.2225		-0.8598	0.1950	
	NS-CSM($\gamma = 0.1$)	-1.4972	0.0672	*	-1.3129	0.0946	*
			Elastic-Net			*Adaptive Elastic-Net*	
CSM	NS-CSM($\gamma = 0.05$)	-0.8241	0.2050		-0.8728	0.1914	
	NS-CSM($\gamma = 0.1$)	-1.5770	0.0574	*	-1.5327	0.0627	*

Table 7: AUC of the benchmark model relative to the network segmented models under the four methods.

We investigate whether the AUC of the network segmented model is significantly different from the benchmark model for the four methods. We applied the DeLong test (DeLong et al., 1988) to investigate the pairwise comparison of the AUC of the benchmark model (i.e., CSM) and that of the NS-CSM for $\gamma = \{0.05, 0.1\}$. We perform these tests under the null-hypotheses that H_0: AUC (CSM) $\geq$ AUC (NS-CSM) and the alternative hypotheses, H_1: AUC (CSM) $<$ AUC (NS-CSM). Table 7 reports the one-sided statistical test of the AUC of the benchmark model relative to the network segmented models. The result of the De Long test shows that while the ROC of CSM is not statistically different from that of NS-CSM($\gamma = 0.05$), the difference between the ROC of NS-CSM($\gamma = 0.1$) and the benchmark (CSM) is statistically significant at 90% confidence level for all four methods.

In conclusion, our proposed factor network approach to credit score modeling presents an efficient framework to analyze the interconnections among the borrowers of a peer to peer platform and provides a way to segment a heterogeneous population into clusters with more homogeneous characteristics. The results show that the lasso logistic model for credit scoring leads to better identification of the significant set of relevant financial characteristic variables, thereby producing a more interpretable model, especially when combined with the segmentation of the population via the factor network-based approach. These empirical results are promising, but certainly not definitive. More research is required to determine whether the observed 'lift' truly is significant rather than just an artifact of random chance or spurious correlation, especially given the fact that these p-values are not calibrated in any way (e.g. Sellke et al., 2001) and Calabrese and Giudici (2015). Further research may include a Bayesian approach, as in Figini and Giudici (2011) and Giudici (2001) We therefore find evidence of a modest improvement in the default predictive performance of our model compared to the conventional approach.

4. Conclusion

This paper improves credit risk management of SMEs engaged in P2P credit services by proposing a factor network-based approach to segment a heterogeneous population into a cluster of homogeneous sub-populations and estimating a credit score model on the clusters using a lasso-type regularization logistic model.

We demonstrate the effectiveness of our approach through empirical applications analyzing the probability of default of over 15000 SMEs involved in P2P lending across Europe. We compare the results from our model with the one obtained with standard single credit score methods. We find evidence that our factor network approach helps to obtain sub-population clusters such that the resulting models associated with these clusters are more parsimonious

than the conventional full population approach, leading to better interpretability and to a modest improved default predictive performance.

References

Ahelegbey D, Giudici P, Hadji-Misheva B. 2019. Latent Factor Models For Credit Scoring in P2P Systems. *Physica A: Statistical Mechanics and its Applications* **522**: 112–121.

Ahelegbey DF, Billio M, Casarin R. 2016a. Bayesian Graphical Models for Structural Vector Autoregressive Processes. *Journal of Applied Econometrics* **31**: 357–386.

Ahelegbey DF, Billio M, Casarin R. 2016b. Sparse Graphical Vector Autoregression: A Bayesian Approach. *Annals of Economics and Statistics* **123/124**: 333–361.

Alam M, Hao C, Carling K. 2010. Review of the literature on credit risk modeling: development of the past 10 years. *Banks and Bank Systems* **5**: 43–60.

Altman EI. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankrupcy. *The Journal of Finance* **23**: 589–609.

Andreeva G, Ansell J, Crook J. 2007. Modelling Profitability Using Survival Combination Scores. *European Journal of Operational Research* **183**: 1537–1549.

Barrios LJS, Andreeva G, Ansell J. 2014. Monetary and Relative Scorecards to Assess Profits in Consumer Revolving Credit. *Journal of the Operational Research Society* **65**: 443–453.

Battiston S, Gatti DD, Gallegati M, Greenwald B, Stiglitz JE. 2012. Liaisons Dangereuses: Increasing Connectivity, Risk Sharing, and Systemic Risk. *Journal of Economic Dynamics and Control* **36**: 1121–1141.

Billio M, Getmansky M, Lo AW, Pelizzon L. 2012. Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors. *Journal of Financial Economics* **104**: 535 – 559.

Calabrese R, Giudici P. 2015. Estimating Bank Default with Generalized Extreme Value Regression Models. *Journal of the Operational Research Society* **66**: 1783–1792.

Carvalho CM, West M. 2007. Dynamic Matrix-Variate Graphical Models. *Bayesian Analysis* **2**: 69–98.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**: 837–845.

Diebold F, Yilmaz K. 2014. On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms. *Journal of Econometrics* **182**: 119–134.

Eichler M. 2007. Granger Causality and Path Diagrams for Multivariate Time Series. *Journal of Econometrics* **137**: 334–353.

Emekter R, Tu Y, Jirasakuldech B, Lu M. 2015. Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending. *Applied Economics* **47**: 54–70.

Figini S, Giudici P. 2011. Statistical merging of rating models. *Journal of the operational research society* **62**: 1067–1074.

Giudici P. 2001. Bayesian data mining, with applications to benchmarking and credit scoring. *Applied Stochastic models in business and industry* **17**: 69–81.

Giudici P, Hadji-Misheva B, Spelta A. 2019. Network-based scoring models to improve credit risk management in peer to peer lending platforms. *Artificial Intelligence in Finance* .

IMF. 2011. Global Financial Stability Report: Grappling with Crisis Legacies. Technical report, World Economic and Financial Services.

Letizia E, Lillo F. 2018. Corporate Payments Networks and Credit Risk Rating. Working paper, https://arxiv.org/abs/1711.07677.

Moghadam R, Viñals J. 2010. Understanding Financial Interconnectedness. Mimeo, International Monetary Fund.

Sellke T, Bayarri M, Berger JO. 2001. Calibration of ρ Values for Testing Precise Null Hypotheses. *The American Statistician* **55**: 62–71.

Serrano-Cinca C, Gutiérrez-Nieto B. 2016. The Use of Profit Scoring as an Alternative to Credit Scoring Systems in Peer-to-Peer Lending. *Decision Support Systems* **89**: 113–122.

Tabachnick BG, Fidell LS, Ullman JB. 2007. *Using Multivariate Statistics*, volume 5. Pearson Boston, MA.

Tibshirani R. 1996. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society. Series B* **58**: 267–288.

Trevor H, Robert T, JH F. 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction.

Viñals J, Tiwari S, Blanchard O. 2012. *The IMF'S Financial Surveillance Strategy*. International Monetary Fund.

Zou H. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American statistical association* **101**: 1418–1429.

Zou H, Hastie T. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the royal statistical society: series B (statistical methodology)* **67**: 301–320.

Zou H, Zhang HH. 2009. On the Adaptive Elastic Net with a Diverging Number of Parameters. *Annals of statistics* **37**: 1733.