

Factorial Network Modeling To Improve P2P Credit Risk Management

Daniel Felix Ahelegbey
Paolo Giudici
Branka Hadji-Misheva

Department of Economics and Management,
University of Pavia, Italy

28 June, 2019
© RegTech Workshop II Frankfurt

- 1 Motivation
- 2 Model Formulation and Estimation
- 3 Empirical Application

- 1 Motivation
- 2 Model Formulation and Estimation
- 3 Empirical Application

- Loans issued by financial institutions, such as banks, to other firms and individuals, are often associated with risks.
- **Credit Risk:** The failure of loan recipients to honor their obligation at the time of maturity.
- This leaves the banks vulnerable and affects their operations.
- To minimize such risk exposures, various methods are used by the credit-issuing institutions to undertake a thorough assessment to classify loan applicants into risky and non-risky customers.
- A conventional individual-level 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.

Motivation

Digital-based Systems (Fin-Tech)

- Digital-based systems are gradually transforming the traditional economic and financial systems
- For example, Peer-to-Peer (P2P) platforms have introduced many opportunities for both lenders and investors and this is redefining the role of traditional intermediaries.
- 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 (1) improves the customer experience, (2) improves the speed of the service and (3) reduce costs to both individual borrowers and lenders as well as small business owners.

Credit Risk Management: Credit Evaluation

Challenges

- Borrower's repayment history & Current capacity to repay (historical financial trends/future cash flow projections)
- Borrower's economic sector status, their expertise and standing in said sector.
- Collect Customer data
 - Business References
 - Financials
 - Credit History
 - Credit Score/FICO Score
- Banks can use collateral, certified accounts, regular reporting to enhance the trust in the borrower.

Credit Risk Management: Credit Evaluation

P2P Challenges

- Characterized by the asymmetry of information:
Lender don't know borrower's credibility as well as borrower does.


Information asymmetry might result in adverse selection - difficult to distinguish healthy and risky credit applicants
- Implementing the traditional system in the online environment which will incur a significant transaction cost
- There is relatively little empirical work in this area due to data accessibility

Why Model Networks?

- To study the pattern of connections in a system¹
 - the **structure** of the pattern of interactions, can have a big effect on the **behavior of the system**.
 - The connections in a social network affect how people learn, form opinions, and gather news, as well as affecting other less obvious phenomena, such as the spread of disease.
 - Unless we know something about the structure of the network, we cannot hope to understand fully how the system works.

Network Models for Credit Scoring

Network models helps to assess borrowers by the company they keep (neighborhood) which can act as signals of credit quality (credibility)

¹Newman, M. (2010), Networks: An Introduction, Oxford university press. 

Objective and Model Scope

We develop a factor-network-based segmentation to improve credit risk assessment for SMEs involved in P2P lending

- The traditional logistic regression for default prediction is enhanced by a clustering exercise through latent factors that are used to construct a similarity matrix
- We threshold the similarity matrix to estimate a network which allows us to segment the heterogeneous population into a connected and non-connected component
- We then treat the two components as two separate clusters and applied a regularized logistic regression to determine the default assessment model.
- Given the predictors may be many, we compared the regularized logistic regression estimated via Lasso, Adaptive Lasso, Elastic Net and Adaptive Elastic Net.

Contributions & Related Works

- 1 Network-based classification methods (Bertini et. al. 2011, InfoSci; Cupertino et. al. 2013, IEEE; Silva et. al. 2012, IEEE; Carneiro et. al. 2012, IEEE; Ahelegbey et. al. 2019, Phys.A).
- 2 Network-based models for credit scoring (West, D. 2000, C&OpR; Ahelegbey et. al. 2019, Phys.A; Giudici et. al. 2019, Frontiers.)
- 3 Credit risk modeling for P2P lending (Andreeva et. al. 2007, EJOR; Barrios et al. (2014), JORS; Emekter et. al. 2015, ApEcon; Guo et. al. 2016, EJOR; Serrano-Cinca and Gutiérrez-Nieto 2016, DecSS)
- 4 Application of factor models for network structure inference (latent factor network models) (Hoff 2008, ANIPS; Sarkar and Dong 2011, Phys.Rev; Durante and Dunson 2014, Biometrika)
- 5 Application of regularized logistic regression methods for credit risk analysis (Hastie et al. 2009; Tibshirani 1996, JRSS(B); Zou 2006, JASA; Zou and Hastie 2005, JRSS(B); Zou and Zhang 2009, AoS)

- 1 Motivation
- 2 Model Formulation and Estimation
- 3 Empirical Application

Logistic Model

- Let Y be a vector of the loan status of n firms
- $Y_i = 1$ if firm- i has defaulted and zero otherwise
- Let $X = \{X_{ij}\}$, $i = 1, \dots, n$, $j = 1, \dots, p$, 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

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

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

Decomposition of Data Matrix by Factors

- X can be considered as points of n -firms in a p -dimensional space
- It can be interpreted as outcomes driven by some underlying firm characteristics.
- X can be expressed as a factor model given by

$$X = FW + \xi \quad (2)$$

where F is $n \times k$ matrix of latent factors, W is $p \times k$ matrix of factor loadings, ξ is $n \times p$ matrix of errors uncorrelated with F .

- The error term ξ is typically assumed to be multivariate normal but F in general case need not be multivariate normal ².
- Lastly, $k < p$ is the number of factors required to summarize the pattern of correlations in the observed data matrix X .

²Tabachnick et. al. 2007, Using Multivariate Statistics.

Factor Network-Based Segmentation

- Let $G \in \{0,1\}^{n \times n}$, where $G_{ij} = G_{ji} = 1 \iff V_i$ is connected to V_j
- Given F , the marginal probability of a link between V_i and V_j is

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

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

- We validate the link between nodes- i and j in G by

$$G_{ij} = 1(\gamma_{ij} > \gamma) \quad (4)$$

where $1(\gamma_{ij} > \gamma)$ 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 .

- We set $\theta = \Phi^{-1}(\frac{2}{n-1})$. We compare $\gamma = \{0.05, 0.1\}$ to capture a sparse but closely connected community.

Regularized Logistic Regression Models

- Lasso:

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

- Adaptive Lasso:

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

- Elastic-Net:

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

- Adaptive Elastic-Net:

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

- 1 Obtain the factors, F , via singular value decomposition (SVD) of X

$$X = UDV = FW + \xi \quad (5)$$

where U and V are orthonormal, and $D = \Lambda^{1/2}$ diagonal matrix of singular values, Λ is diagonal matrix of eigenvalues of $X'X$ and XX' .

- 2 Set k , number of factors that account for approx 95% variation in X
- 3 Estimate the network \hat{G} determined by

$$\hat{G}_{ij} = 1 \left(\Phi [\theta + (FF')_{ij}] > \gamma \right) \quad (6)$$

- 4 Estimate regularized logistic regression on connected and non-connected clusters.
- 5 Use ten-fold cross-validation to select λ , we chose $\lambda.min$ over $\lambda.1se$.
- 6 Set $\alpha = 0.5$, $v = 2$, and $\hat{\beta}_j$ as the ridge regression estimate.

- 1 Motivation
- 2 Model Formulation and Estimation
- 3 Empirical Application

- 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.
- Total number of institutions (%): Active (13413 - 89.15%) and Defaulted (1632 - 10.85%)
- Data Source: European External Credit Assessment Institution (ECAI)

Application

Data: Summary

Var	Formula (Description)	Active	Defaulted
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

Table: Description of the financial ratios with summary of mean statistics.



Results

Latent Factor Network

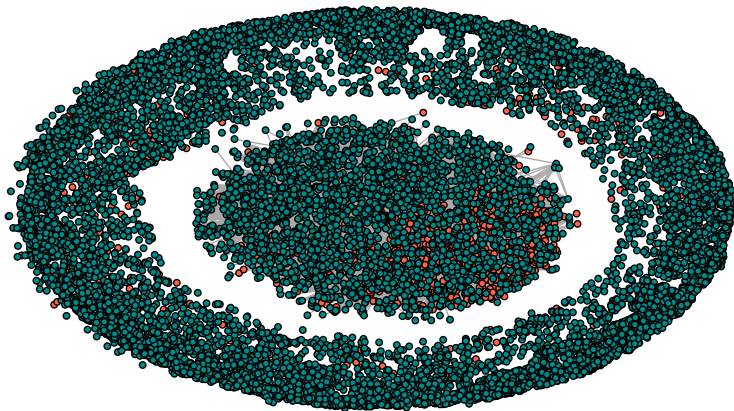


Figure: A graphical representation of the estimated factor network.

Results

Eigen Decomposition of Connected Component

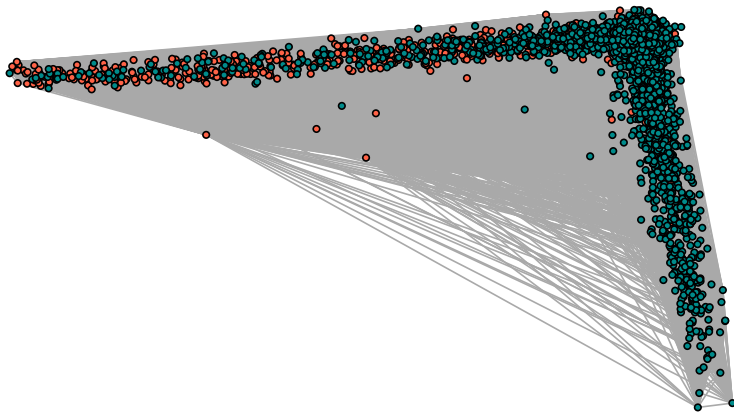


Figure: Representation of the latent positions of connected components.

- Through the eigen decomposition of the network of connected SME's, we can assess the neighborhood of the firms which act as a signal of credit quality (credibility)

Results

Network-based Classification and Default Status

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

- Majority of the healthy (non-defaulted) credit applicants fall within the non-connected component
- Majority of the risky (defaulted) credit applicants fall within the connected component

Results

Comparing Regularized Logistic Regression Models

	CSM	(C)	(NC)	CSM	(C)	(NC)	CSM	(C)	(NC)	CSM	(C)	(NC)
	<i>Lasso</i>			<i>Adaptive Lasso</i>			<i>Elastic-Net</i>			<i>Adaptive Elastic-Net</i>		
V1	0.042	0.015	0.091	.	.	.	0.043	0.014	0.101	.	.	.
V2	0.018	0.072	.	.	0.073	.	0.019	0.071	.	.	0.076	.
V3	-0.555	-0.239	-0.783	-0.631	-0.251	-1.028	-0.556	-0.237	-0.712	-0.63	-0.25	-1.018
V4	-0.246	-0.171	-0.418	-0.203	-0.178	-0.482	-0.273	-0.171	-0.435	-0.202	-0.179	-0.485
V5	0.009	0.036	.	-0.002	.	.	.
V6	0.04	0.007	0.041	0.007
V7	0.19	0.173	0.209	0.154
V8	-0.347	-0.342	.	-0.35	-0.35	.	-0.345	-0.339	.	-0.351	-0.35	.
V9	-0.029	0.026	-0.348	.	.	-0.387	-0.028	0.025	-0.357	.	.	-0.387
V10	0	-0.056
V11	0.042	0.035	0.004	.	.	.	0.098	0.035	0.026	.	.	.
V12	0.075	0.028	0.021	.	.	.	0.08	0.026	0.027	.	.	.
V13	-0.098	-0.185	-0.115	-0.168	-0.009	.	.	.
V14	-0.14	-0.005	.	-0.133	.	.	-0.151	-0.01	0	-0.133	.	.
V15	0.001	-0.095	.	.	-0.099	.	0.012	-0.092	.	.	-0.101	.
V16	0.134	0.323	.	.	0.251	.	0.145	0.314	.	.	0.253	.
V17	0.235	0.197	0.214	0.309	0.214	0.302	0.227	0.201	0.215	0.308	0.211	0.286
V18	0.074	0.163	.	.	0.17	.	0.077	0.163	.	.	0.173	.
V19	0.185	-0.009	0.167	0.194	.	.	0.186	-0.008	0.168	0.194	.	0.016
V20	0.012	-0.024	0.129	.	.	0.179	0.018	-0.025	0.162	.	.	0.185
V21	0.223	0.187	0.309	0.232	0.179	0.351	0.223	0.186	0.31	0.232	0.178	0.344
V22	0.1	.	0.109	0.071	.	0.042	0.098	.	0.121	0.072	.	0.051
V23	-0.169	-0.125	-0.373	-0.156	-0.119	-0.527	-0.171	-0.125	-0.37	-0.155	-0.125	-0.515
V24	0.007	-0.073	0.011	.	-0.038	.	0.007	-0.073	0.02	.	-0.058	.

Results

Comparing Regularized Logistic Regression Models

	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: Number of selected default predictors.

- Enet (less parsimonious), Lasso, AEnet & ALasso (most parsimonious)
- The network-based segmentation framework is therefore more parsimonious than the benchmark full population credit score model, and this helps in interpretability.

Results

Comparing Default Predicting Accuracy

	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: Comparing area under the ROC curve (AUC) of the regularized methods.

Results

Comparing Default Predicting Accuracy

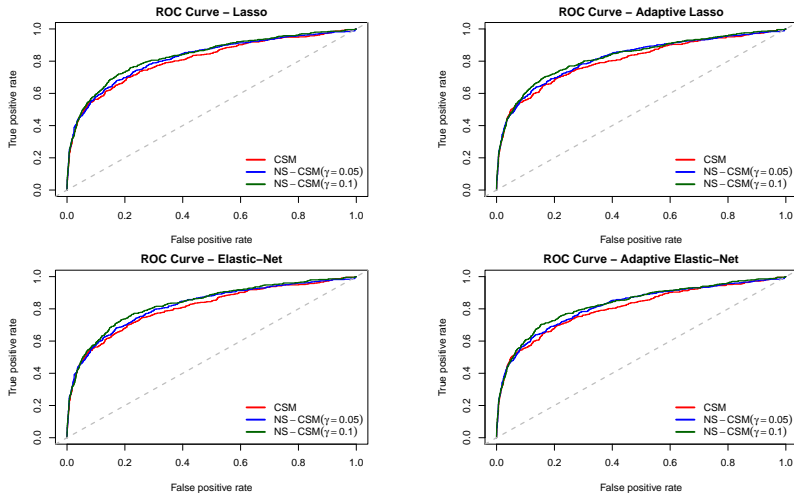


Figure: ROC curves of the regularized methods.

Results

Comparing ROC Curves

Using DeLong test³

		Statistic	Pvalue	Statistic	Pvalue
		<i>Lasso</i>		<i>Adaptive Lasso</i>	
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 *
		<i>Elastic-Net</i>		<i>Adaptive Elastic-Net</i>	
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: AUC of the benchmark model relative to the network segmented models.

- ROC of CSM not statistically different from 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

³DeLong et. al. 1988, Biometrics

- We develop a credit risk assessment approach for SME's and motivates this model to be usable in a peer-to-peer lending.
- The traditional logistic regression for default prediction is enhanced by a clustering exercise through latent factors that are used to construct a network via thresholding of a similarity matrix
- The connected and non-connected components are treated as two separate clusters and regularized logistic regression approaches are applied to determine the default assessment model.
- Evidence that Elastic-net are less parsimonious, followed by Lasso. Adaptive Elastic-net & Adaptive Lasso are most parsimonious
- Evidence that our approach produce a more parsimonious and interpretable model with a modest improvement in the default predictive performance.

Thank you !
for your attention

Research supported by

- Funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825215 (Topic: ICT-35-2018 Type of action: CSA);

Q&A?