Acta Technica *62* (2017), No. 5B, 341–362

The application of principal component analysis and logit model in the measurement of industry credit risk - Taking China as an example

Tianming Cai[1]

Abstract. In order to accurately measure the risk of default in different industries,this paper constructs the index system of measuring industry risk from industry characteristic risk factors, industry operating conditions and financial risk.Then the principal component analysis (PCA) is used to extract main influence factors.Then the Logit model is built to measure credit risk of Chinese industry and the model passes validity test.The results show that default risk of the leather, furs, down and related products and food processing industry is higher.The default risk of electric power industry, power industry, telecommunications, petroleum and petrochemical industry is lower.These results can adjust the loans structure for the banking, provides the theory basis for lower credit risk.

Key words. Industry credit risk, Principal component analysis, Logit model, Index construction.

1. Introduction

Credit risk is a long-term and ubiquitous risk in the financial market.A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments[1].When the economy is upgoing,the credit will create new employment opportunities and accelerate the economic growth.During the recession, however, high credit risk may affect the stability of domestic financial and economic development, even affect the stability of the global economy.From the United States "subprime crisis" in 2007 to Iceland "bankruptcy", from Dubai to Greece's debt crisis in 2009, the damage of credit risk mismanagement is obvious[2,3].

Because of the credit risk mismanagement the original state-owned commercial banks of China had the huge non-performing assets.In 1998, 1.4 trillion yuan

[1]Workshop 1 - Research Center for Finance Computing, Ningbo Dahongying University, Ningbo, China;email:nbctm@163.com

non-performing loans of the four state-owned commercial banks were hived off into four financial asset management companies.In 2003 the Bank of China and China Construction Bank started shareholding system reform.In 2009 Agricultural Bank of China finished shareholding system reform.The state-owned commercial banks through the capital injection, equity restructuring, the introduction of strategic investors and IPO, built up modern banking system after several years, and non-performing assets ratio reduced to the international level.According to the annual report of the China banking regulatory commission (CBRC), by the end of 2016, the loan balance of the national banking financial institutions was as high as 106.6 trillion yuan and non-performing loan balance wan 1.96 trillion yuan.This shows that the credit risk management of commercial bank is still the important content in China.Accurate measuring credit risk, however, is the premise and basis of credit risk management for commercial banks.No accurate measuring credit risk, the bank loans or investment will not be able to make the right decisions, and it is impossible to formulate a reasonable price on lending and other conditions of contract,and it is more impossible to take appropriate measures to control or shifting risk after lending[4].

In terms of the credit risk measure, traditional methods of measuring credit risk have the experts method, the rating method and the credit score method. The experts method is one of the most traditional credit risk analysis method.The most famous experts method is "5c" factor analysis method, it is analyzed by loan officer who have rich experience according to the Character,Capital, Capacity, Collateral and Cycle Conditions of the borrower, and then suitable weight is given to these five indexes to finally determine the lending decisions.The office of the comptroller of the currency (OCC) developed the original rating method.The rating method is mainly used by the regulators and bankers of United States and other countries to assess the adequacy of loan loss provisions.This method divides loan portfolios into normal, focus on, secondary, suspicious, and loss five classes, different ratings are endowed with different loss reserve requirements.After calculating the loss reserves and adding up, it gets the capital that commercial Banks need to prepare for risk.Credit score method mainly uses mathematical statistics method to establish regression model, calculates the credit score or default probability of evaluation objects to determine the size of the credit risk.Credit score method is the most types, the most widely used method of credit risk measurement.Credit rating method mainly includes multiple discriminant analysis,Logit model, Probit model and neural network method, etc[5].

The present method is more on enterprise and individual, traditional evaluation methods tend to emphasize the risk of enterprise itself, and don't fully consider the function of the industry to which the enterprise belongs [6,7].Under the environment of the deepening marketization, the influence of industry development to the enterprise is becoming more and more obvious.Therefore, the industry credit problems should be paid greater attention.

There is no unified rule to current industry credit risk limits. According to some literature analysis [8,9,10], industry credit risk needs to be on the basis of industry measuring credit risk.So it is necessary to research on the industry measuring credit risk. It can provide decision-making basis for commercial bank's credit in-

vestment decisions, reduce the credit risk in the investment decision-making process of commercial bank credit.

2. measuring Industry credit risk Based on the PCA - Logit model

2.1. The construction of index system

Industry is different from normal enterprise risk, it involves more factors[11].It can't draw lessons from the practice of the normal enterprise risk,but should consider more factors determine an industry's average probability of default risk.Because industry risk including industry environment, industry operating conditions,industry financial conditions.Many Banks in China only do single static research of industry risk,in addition, on the basis of many factors, only give qualitative analysis or give a analysis framework. There is no specific research, the industry risk factors are not quantitative.Based on this,this paper attempts to give a analysis of all aspects of the industry, and build index system of measuring industry credit risk.

(1) Industry characteristics of the risk factors

Industry characteristics of the risk factors are mainly analyzed from industry development degree, the degree of monopoly industry, industry dependence of its upstream and downstream products and market development potential, etc.And then judge the change of industry market competition and the development of the industry in the future prospects and trends.It can provide the basis for bank credit decisions.

Industry life cycle(x1)

Industry life cycle index is mainly used to reflect industry development degree.Industry development general experience start-up stage, growth stage, mature stage and decline stage.This qualitative index can be incorporated into the model after quantitative.Quantitative method is investigation and expert scoring.Specific criteria is the higher the score the smaller the risk, according to the 4 grade scale. For industry recession is obviously at highest risk and minimal score. In the late development period and late mature period, industry has the most potential for development.So risk of bank loans is the minimum, the highest score should be.Specific scoring criteria is shown in table 1.

Industry monopoly(x2)

Industry monopoly factor is mainly used to reflect industry economic structure characteristics.Depending on the economic structure, industry basically can be divided into four types of markets: perfect competition, imperfect competition and monopolistic competition, oligopoly and monopoly.

Industry dependence (x3)

Industry dependence refers to the target industry operating conditions and the relevance of the upstream and downstream industries.The dependence of the target industry of its upstream and downstream industry, the greater the potential risks, lending the higher.

Alternative industry products (x4)

Alternative industry products to some extent reflects the development prospect of the industry. If the industry product has many substitute products from other industries, then the industry will face greater competitive pressure, especially when substitute products have a comparative advantage in terms of price and performance, the potential risk of target industry might be at a higher level.

Industry economic cycle (x5)

Industry economic cycle mainly refers the relational degree of development of the industry and changes in the economic cycle. According to this relational degree, the industry is divided into: a. Growth industries. The motion state of growth industries has nothing to do with the period and amplitude of total level of economic activity.b. Cycle type industry. The motion state of cycle type industry directly relates to the economic cycle.c. Hysteresis cycle industry.The motion state of hysteresis cycle industry also relates to the economic cycle, but may be imperceptible on the surface. This is mainly because this kind of industry will not immediately impact by the change of the economic cycle, but will indirectly in quite a long time is affected.d. Defensive industry.These industry motion exists because its product demand is relatively stable, not affected by economic cycle in the recession stage. Also, such index belongs to the qualitative class, need to quantify.

In conclusion, according to the higher the score, the smaller the risk criteria, the quantitative criteria of the above five qualitative index as shown in table 1.On the basis of the investigation, combined with expert inquiry, according to the scoring method, finally through weighted will receive a score of each index.These scores as the original data of the index analyse with the back of the quantitative index.

Table 1. Quantitative industry characteristic factors

	1	2	3	4
Industry life cycle(x1)	Decline phase	Start-up stage	Late period of early maturity	Development in the late early maturity
Industry monopoly (x2)	Perfect competition	Imperfect competition and monopoly competition	Oligopoly	Perfect monopoly
Industry dependence (x3)	Upstream and downstream industry concentration degree is high	Upstream and downstream industry concentration degree is a bit high	Upstream and downstream industry concentration degree is a bit low	Upstream and downstream industry concentration degree is low
Alternative industry products (x4)	High	A bit high	A bit low	Low
Industry economic cycle (x5)	Cycloid type	The hysteresis cycle type	Defensive	Growth

(2)Industry operating conditions and financial risk

Considering the data availability and comprehensiveness, industry operating conditions and financial risks are mainly from the industry's profitability, development ability, capital liquidity and industry operation efficiency to reflect, the meaning of pecific index is as follows:

A.Industry profitability

Industry profitability is one of the important factors affect the industry average reimbursement ability, and the index of reflecting the industry average profit ability including industry total return on assets, return on sales and profit growth rate, percentage of loss-incurring enterprises, industry losses, etc.The specific meaning and calculation for each index is as follows.

Return on assets (x6) : Return on assets refers to the whole industry in a certain period of profit and the ratio of average assets.It is the most important index of industry profitability, fully embodies the ability of its own capital to earn profits.

Return on sales (x7) = profit/sales: It refers to the enterprises in the industry a period of time sales profit with the ratio of sales revenue.It shows that the industry average revenue per unit can bring how many sales profit.It reflects the average enterprise main business profitability in the industry. It is the main index in the management benefit evaluation industry.

Industry percentage of loss-incurring enterprises (x8) = number of unprofitable enterprises in the industry/total number of enterprise in the industry: Industry percentage of loss-incurring enterprises is an important index of degree of risk.It reflects the operating condition of the unprofitable enterprises in the target industry. Industry percentage of loss-incurring enterprises is equal to the number of unprofitable

enterprises in the industry and the ratio of the number of all enterprises in the industry.

Loss of industry (x9) = total loss of industry/total net profit of industry: Loss of industry measures the degree of enterprise operating losses in the industry.It expressed as the absolute value of the total losses divided by the sum of the absolute value of total earnings and the absolute value of total lossis in in the industry.This value is between 0 and 1.

B.Industry development ability

Rate of capital accumulation (x10) : Industry rate of capital accumulation is refers to the enterprises in the industry this year at the beginning of owner's equity increase and the ratio of owner's equity.Rate of capital accumulation shows the capital accumulation ability, It is an important index of evaluation target industry development potential.The higher the rate of capital accumulation, the stronger the capital preservation, the greater the industry against risk and ability of sustainable development.

Product sales revenue growth (x11) : It shows product marketing development ability.

Profit growth rate (x12) : Profit growth rate refers to the final profit total profit growth rate than the same period last year.It represents the industry in the growth of profit, reflects the degree of the development of the industry.

C. Industry capital liquidity

Liquidity is a comprehensive index of asset value cashability of the industry. So liquidity is a reflection of assets liquidation ability to repay loans, to some extent also reflects the industry's solvency.It mainly uses current asset turnover and liabilities rate index to reflect.

Current assets turnover ratio (number) (x13) :It is a period of time net sales revenue and the industry liquidity ratio between the average occupancy.Its computation formula is Current assets turnover ratio (number)= sales net income/average balance current asset.

Liabilities rate (x14):It refers to a certain period of time the growth of the industry in the process of operating debt capital.It can indirectly reflect the industry's capital liquidity.

D. Industry operating capacity

Cost margins (x15):It reflects economic benefit of the production costs invested by the industry, but also reflects economic benefits of the industry reduce the cost.Its calculation formula is: cost margins = total profit/total cost.Among them: the total cost is sum of the product cost of sales, marketing fees, management fees and financial costs.

Capital preservation (x16) = the end of report period owner's equity/the end of the same period of last year owner's equity.This index reflecs the industry changes in the average net assets.It is the concentrated reflection of industry business development ability.

The above indexes can be obtained directly statistics, part indexes need to be calculated on the basis of the original data.

2.2. The reason for choosing PCA - Logit model

The risk measure of this paper is aimed at industry risk.For the industry risk measure mainly is the average probability of default, the reason of using Logit model to measure the probability of the risk of industry default is mainly based on the following consideration:

First of all, industry risk indexes have more with high dimension characteristic, suitable for Logit model of high-dimensional data processing.The disadvantage of Logit model is difficult to measure in the qualitative indexes. So before application, for some qualitative indexes of risk,through expert evaluation and investigation method to quantify to overcome its shortcomings.

Secondly, the general multiple discriminant analysis method depends on normal distribution assumption, and Logit model uses maximum likelihood estimation method for parameter estimation, does not require sample data obey the normally distributed.

Third, the probability of default (PD) measurement, the result is limited to between 0 and 1, so the general linear model is not good.Although, in general linear regression model is very popular in the quantitative analysis of the statistical analysis method, but considering the calculation model of PD, as the dependent variable is a binary classification variable (such as: normal or default), rather than a continuous variable, so for the binary classification dependent variable analysis need to use nonlinear function, and Logit distribution is the most commonly used functions.

2.3. Construction Logit model

1. The model form

Logit model actually is the development of general multivariate linear regression model, the z value using linear discriminant model is just an abstract concept, can only be used for judging not to be intuitively explained.Logit model solved the problem.After construction the industry index, it can calculate for a period of time the probability of default through the Logit model. If it is concluded that the probability is greater than the set point, it can consider that the industry has a high credit risk.

Logit model is the basic binary classification variables (such as success and failure, pro and con) or multiple classification variables (such as senior, secondary, junior). The initial condition of observation the default is the default or no default in this paper , and this binary classification variable is Logit model assumptions. This binary classification variables is respectively with 0 and 1 representing not default and default.Because the default rate of the industry is hidden variable, the processing method is corresponding to an observable z-points variable which will be subject to non-performing assets ratio.Non-performing assets ratio is the ratio of non-performing assets to total assets.Non-performing assets ratio emphatically reflects the quality of enterprise assets from the assets which the enterprise can't normal cycle for income.It reveals the problems of the enterprise on the asset management and use. According to the the commercial bank risk regulation core indicators which

is issued by China banking regulatory commission in 2005, in this paper Logit model dependent variable is defined as follows:

$$Y_l = \begin{cases} 1 & \text{non} - \text{performing loan ratio} \geq 4\% \\ 0 & \text{non} - \text{performing loan ratio} < 4\% \end{cases} \tag{1}$$

The Logit model assumes that probability events obey the standard logistic cumulative probability distribution:

$$Y_l^* = \frac{1}{1 + e^{-X_i B}}$$

Among them,$X_i = (X_0, X_1, ..., X_m)$is principal component factors which are extracted earlier in this paper,$B = (\beta_0, \beta_1, ..., \beta_m)$is parameters to be estimated.Assume that P is the industry average probability of default, 1 - P is the industry average probability of loan repayment on time, then

$$P_i = E(Y_i = 1|X_i) = \frac{1}{1 + e^{-X_i B}}$$

Converte to linear function:$\frac{p_i}{1-p_i} = e^{X_i B}$

Take the logarithm and the linear function is:

$$LogitP = \ln \frac{p}{1-p} = \beta_0 + \beta_1 y_1 + ... + \beta_m y_m \tag{2}$$

You can get

$$P = \frac{1}{1 + e^{-\beta_0 - \beta_1 y_1 - ... - \beta_m y_m}} \tag{3}$$

Obviously,$P \in [-1, 1]$,(3)is the Logit regression model.

2.Parameter estimation

Set default sample for high-risk industry$Y_1^A, Y_2^A, ..., Y_h^A$,no default samples for low risk industry$Y_1^B, Y_2^B, ..., Y_h^B$,then Likelihood function is:

$$L = \prod_{i=1}^{h} [1 - \frac{1}{1 + e^{-\beta_0 - \beta_1 y_1^A - ... - \beta_m y_m^A}}] \prod_{i=1}^{h} [1 - \frac{1}{1 + e^{-\beta_0 - \beta_1 y_1^B - ... - \beta_m y_m^B}}] \tag{4}$$

Take the log of above formula, and calculate maximum likelihood estimator$B = (\beta_0, \beta_1, ..., \beta_m)$using the maximum likelihood estimation method,that is taking the logarithm likelihood function (lnL) to maximize.While estimating the B,to B calculates partial differential with (lnL) and sets the result to 0, you can get the estimated parameters$(\beta_0, \beta_1, ..., \beta_m)$.

On the basis of Logit model, and as a result of the Logit model defects, this paper also selects the PCA (Principal Component Analysis)method,which is mainly based on the following consideration: under normal circumstances, the index requirements of PCA method has the high correlation, which can get a better evaluation results, and between the principal components are not related.If there is a strong relationship

between some indexes, it can produce the multicollinearity, in this case, only keep one of these indexes, and as a representative of the other indexes;Otherwise, the multicollinearity will reduce the effectiveness of the Logit results. As the main component in the PCA is unrelated, so principal component and regression analysis help Logit model of solving multicollinearity problem, make the Logit model more effective.

3. measuring credit risk of Industry using PCA-Logit model

3.1. Sample selection and data sources

1. Sample selection and grouping

This paper choose the data which can reflect the true state of the industry as the research object. The industry data is from 101 industries which are under statistical caliber of the state-owned assets supervision and administration, including electricity industry, telecommunications, agriculture, fishing, oil processing and coking industry, etc.

Samples were divided into two groups, the samples of 2012 are training samples, the samples of 2013 are predicting samples. It is generally recognized that the enterprise risk is based on non-performing loan ratio. Because here is the study of the industry, but the industry average non-performing loan ratio is difficult to statistic in China due to the defect of historical data, so this study with non-performing asset ratio as the boundary between the high risk and low risk enterprises. According to the industry statistics from IFinD database,the non-performing asset ratio data is the foundation. According to the function(1) the training sample and predicting samples can be divided into high and low risk industry.

2.Variable definitions and data sources

According to the index system built in previous section, to facilitate analysis, respectively define variable indexes by label xl -x16.Industry environmental factor corresponding indexes, according to the previous analysis only have the effect of correction of results, so there is no definition.

The 16 indexes data of 101 industry sources is:the industry statistics from IFinD database, financial yearbook, statistical bulletin and statistics website,part as the original statistical data, part of the calculation.

3.2. Principal component analysis to extract the Logit model independent variable

In order to avoid the correlation of these indexes variables and the effects of multicollinearity, principal component analysis is used.

1.Introduction of the principal component analysis

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

Consider a data matrix, X, with column-wise zero empirical mean (the sample mean of each column has been shifted to zero), where each of the n rows represents a different repetition of the experiment, and each of the p columns gives a particular kind of feature.

Mathematically, the transformation is defined by a set of p-dimensional vectors of weights or loadings $w_{(k)} = (w_1, ..., w_p)_{(k)}$that map each row vector$x_{(i)}$of $X$to a new vector of principal component scores$t_{(i)} = (t_1, ..., t_m)_{(i)}$,given by

$t_{k(i)} = x_{(i)} \cdot w_{(k)}$ for $i = 1, ..., n; k = 1, ..., m$

in such a way that the individual variables$t_1, ..., t_m$of $t$considered over the data set successively inherit the maximum possible variance from$x$, with each loading vector w constrained to be a unit vector.

In order to maximize variance, the first loading vector$w_{(1)}$thus has to satisfy

Equivalently, writing this in matrix form gives

Since$w_{(1)}$has been defined to be a unit vector, it equivalently also satisfies

The quantity to be maximised can be recognised as a Rayleigh quotient. A standard result for a positive semidefinite matrix such as $X^T X$ is that the quotient's maximum possible value is the largest eigenvalue of the matrix, which occurs when $w$is the corresponding eigenvector.

With$w_{(1)}$ found, the first principal component of a data vector $x_{(i)}$can then be given as a score $t_{1(i)} = x_{(i)} \cdot w_{(i)}$in the transformed co-ordinates, or as the corresponding vector in the original variables, $\{x_{(i)} \cdot w_{(1)}\}w_{(1)}$.

The kth component can be found by subtracting the first$k - 1$principal components from X:

and then finding the loading vector which extracts the maximum variance from this new data matrix

It turns out that this gives the remaining eigenvectors of $X^T X$, with the maximum values for the quantity in brackets given by their corresponding eigenvalues. Thus the loading vectors are eigenvectors of $X^T X$.

The kth principal component of a data vector$x_{(i)}$can therefore be given as a score $t_{k(i)} = x_{(i)} \cdot w_{(i)}$ in the transformed co-ordinates, or as the corresponding vector in the space of the original variables, $\{x_{(i)} \cdot w_{(k)}\}w_{(k)}$, where $w_{(k)}$ is the kth eigenvector of $X^T X$.

2. The correlation coefficient calculation

First this paper does KMO and Bartlett's test, the results are shown in the following table 2.

Table 2. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.804
Bartlett's Test of Sphericity	Approx. Chi-Square	1655.291
	df	276
	Sig.	.000

From table 2 shows that Bartlett's test statistics of the observed value is 1655.291, the corresponding probability P is close to 0, less than the significance level (P = 0.05).So refuse to null hypothesis,the correlation coefficient matrix with the unit matrix have significant difference;At the same time, KMO value is $0.804 > 0.5$, so the principal component analysis was suitable for carried out on the input index.

3.Extracting principal component

According to the results of KMO and Bartlett's test above,principal components were extracted with principal component analysis (PCA).The initial solution of principal component analysis was got, namely common data of all the input index, which is shown in table 3.Communalities depicts how principal component expain the index information.It can be used to evaluation how index of information loss, and is the important basis to measure the effect of principal component analysis.Communalities is more close to 1, the more principal component explain the index variance,the less loss of information;On the contrary, the opposite.

Table 3. Communalities

Factor	Initial	Extraction	Factor	Initial	Initial
x1	1.000	.802	x9	1.000	.780
x2	1.000	.779	x10	1.000	.797
x3	1.000	.739	x11	1.000	.788
x4	1.000	.807	x12	1.000	.742
x5	1.000	.803	x13	1.000	.697
x6	1.000	.865	x14	1.000	.506
x7	1.000	.833	x15	1.000	.791
x8	1.000	.653	x16	1.000	.860
Extraction Method: Principal Component Analysis.					

The third column of table 3 are communalities when extracting the characteristic values acccording to the specified conditions.Part of the communalities of input index are lower, such as x14 is only 0.506.In order to better reflect the correlation between indexes,we consider to remove the variables of lower communalities, and then do principal component analysis again, finally get the following final result.

Table 4. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.751
Bartlett's Test of Sphericity	Approx. Chi-Square	753.305
	df	45
	Sig.	.000

Table 4 shows that Bartlett's Test of Sphericity value is 753.305, the corresponding probability P is close to 0, less than the significance level ($P = 0.05$), so refuse to null hypothesis, accept that the correlation coefficient matrix with the unit matrix have significant difference;At the same time, KMO value is $0.751 > 0.5$, so it is suitable for doing the principal component analysis for input indexes.

Table 5. Communalities

Factor	Initial	Extraction	Factor	Initial	Initial
x1	1.000	.808	x10	1.000	.898
x3	1.000	.820	x11	1.000	.893
x5	1.000	.839	x12	1.000	.724
x6	1.000	.854	x15	1.000	.723
x7	1.000	.802	x16	1.000	.754
Extraction Method: Principal Component Analysis.					

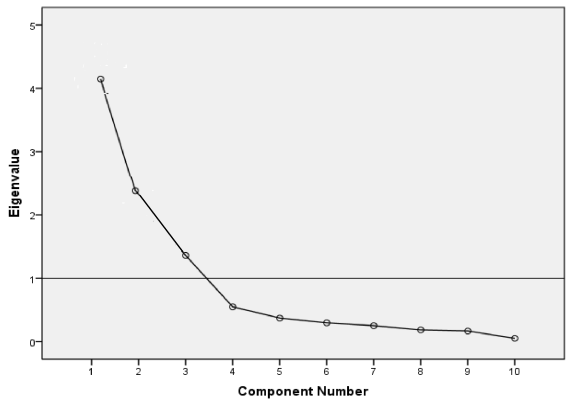

Fig. 1. Scree Plot

From Figure 1, it can be seen that the scattered points of the first 3 factors are on the steep slope, and the latter 7 factors are all less than 1, so only the first 3

common factors are considered.

The communalities value when extracting the characteristic values from table 6 show that the principal component can reflect most of the input indexes information(80%), only little information loss, principal component analysis effect is good. Then do analysis of the characteristic value and total cumulative variance contribution rate, and the results are shown in table 6 below.

Table 6. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	%of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.246	42.461	42.461	4.246	42.461	42.461	3.412	34.119	34.119
2	2.344	23.436	65.897	2.344	23.436	65.897	2.628	26.278	60.397
3	1.426	14.256	80.154	1.426	14.256	80.154	1.976	19.757	80.154
4	.611	6.111	86.264						
5	.408	4.078	90.342						
6	.319	3.188	93.530						
7	.283	2.827	96.358						
8	.173	1.732	98.090						
9	.142	1.417	99.507						
10	.049	.493	100.000						
Extraction Method: Principal Component Analysis.									

From table 6,the cumulative contribution rate of the first three principal components is 80.154%, the first three principal component can reflect the 80.154% information of the original index.According to the principle of cumulative contribution rate more than 80% and eigenvalues greater than 1,so the extraction of three principal component is appropriate.

In order to naming principal components and having named explanatory, in this paper,the principal component load matrix of the indexes is maximum variance rotated.The three principal components load matrix of the indexes are shown in the table 7.

Table 7. Rotated Component Matrix

Factor	Component		
	1	2	3
x10	.905		.270
x11	.894		.303
x12	.865	.244	
x16	.814	.108	-.221
x1	-.103	.899	
x3	.342	.855	
x5	.426	.733	.288
x6	.227	-.191	.798
x7	-.145	.494	.758
x15		.439	.678

According to correlation between the principal component of principal component load matrix and the original,the original index x10, x11, x12, x16 have a higher load on the first principal component, so the first principal component reflects these basic indexes of information;x1, x3, x5 have a higher load on the second principal component, so the second principal component reflects these basic indexes information;x6, x7, x15 have a higher load on the third principal component.According to the information of indexes meaning, three principal components, respectively, represents the industry's development capacity, industry's development of the cyclical and the profitability of the industry, respectively for X1, X2, X3.

4. Calculate principal component scores

According to table 7, the eigenvectors corresponding eigenvalue of the three principal component are calculated, followed by,$\lambda_1\lambda_2\lambda_3$ which are shown in table 8 below.

Table 8. Component score coefficient matrix

Factor	Component			Factor	Component		
	λ_1	λ_2	λ_3		λ_1	λ_2	λ_3
x10	.260	.052	-.112	x3	.262	-.080	.113
x11	-.102	.410	-.124	x5	.275	-.125	.111
x12	-.128	.109	.375	x6	.040	-.236	.490
x16	.041	.348	-.113	x7	.265	.025	-.196
x1	.064	.253	.024	x15	-.058	.084	.325

The principal component scores formula is

$$F_i = \lambda_i Z x_{ij}, i = 1, 2, 3; j = 1, ...101$$

which $Zx_{ij}$is standardized data.After the characteristic vector and standardized values substitute into the formula above, score vectors of 101 industries in the three principal component are shown in table 9.

Table 9. The principal component score vectors of various industries

The name of the industry	F1	F2	F3
Salt industry	0.47812	0.19209	-0.98379
Animal husbandry	-0.22511	-0.22084	0.54376
The ship manufacturing	-0.7515	-1.31281	-1.06072
Electrical equipment manufacturing	-1.35852	-0.74213	0.53593
Electric manufacturing	1.03556	-0.77304	-0.79753
Electric power industry	-0.86726	-0.09632	-0.99882
The power supply industry	-1.67428	3.84199	1.86986
Electric power industry	-0.26126	1.73712	-0.74474
Electrical machinery and equipment manufacturing	0.85543	-0.99812	-0.39257
Telecommunications industry	0.29387	4.11972	-2.10123
The electronics industry	-0.86726	-0.09632	-0.99882
The electronic computer manufacturing	-0.53666	0.2041	-0.8148
Electronic component manufacturing industry	-0.60105	0.60533	2.06854
Textile industry	-1.05338	-0.23392	0.53046
Fertilizer manufacturing	0.06896	0.66041	0.24707
Clothing, shoes, hats manufacturing	-0.38856	-0.45619	0.54365
Arts and crafts and other products	-0.13397	-0.35297	-0.70477
Radio and television equipment manufacturing	-0.75565	-0.81328	-0.50447
Boiler and prime mover manufacturing	0.59751	-1.0489	-0.58366
Ferrous Metals Mining and Dressing	-0.3398	0.00139	0.03631

4. The model validation

First test goodness of fit of the model,which mainly using - 2 logarithm likelihood values (- 2 ll).The likelihood value range is between 0 and 1, the scope of the logarithm is between negative infinity and 0.

Second,test the determine effect of the model,testing the training sample and predicting sample, respectively.This involves the problem of point selection, which will affect to the determine effect of the model.Different from linear discriminant model, Logit model doesn't have theoretical threshold, the threshold need to select according to the research target.Assume that threshold is 0.5, it means the industry's default probability calculated by Logit model is greater than or equal to 0.5, and the industry will be judged to be high risk type, otherwise the industry will be judged to be low risk type, in order to test the predicting effect and misjudgement. For wrongly types, there are two types of statistics of misjudgment, the first kind of error is called the "true", the second type of error is called the "false". In credit risk assessment, also use these two types of error measure discriminant model prediction accuracy,which the first error is high credit risk type mistaken for low credit risk types, the second type of error is low risk type mistaken for high credit risk types.General misjudgment rate is below 30% as well.

4.1. Logit regression parameter estimation and the validity test

1. Parameter estimation

Put the above three principal component Xl, X2, X3 into the Logit model to estimate parameters.The Logit model parameters calculated by SPSS are shown in the table 10.

Table 10. Variables in the equation

		B	S.E,	Wals	df	Sig.	Exp (B)
step 1	X1	-.196	.243	.655	1	.418	.822
	X2	-.862	.400	4.656	1	.031	.422
	X3	.226	.262	.742	1	.389	1.253
	constant	-1.175	.265	19.600	1	.000	.309

Among B is variable regression coefficient;S.E is standard error;Wals is used to determine whether a variable is included in the model, and the greater Wals value,the stronger the correlation between dependent and independent variables;Sig is significance level, this value is as small as possible.Exp (B) is occurrence rate, if this value is greater than l, means that the independent variable has a positive effect of probability of occurrence;If this value is less than 1, the independent variable has a negative effect to the event probability.

According to table 11,Logit regression function is

$$LogitP = (-1.175 - 0.196X_1 - 0.862X_2 + 0.226X_3)$$

so,$P = \frac{1}{1+e^{(-1.175-0.196X_1-0.862X_2+0.226X_3)}}$ (5)

2. The validity test

Verification must be conducted after model parameter estimation.First examine the suitability of the model, by observing the output you can see that the -2ll difference is 109.621 and is greater than the corresponding chi-square threshold, which means the model is appropriate.Secondly test its effect,which involves a tipping point selection problem and will influence determine effect of the model.Because in theory Logit model doesn't have the optimal point of division,the point selection depends on the model of the user's specific goals.Traditionally that P value point of division is 0.4.Put predicted sample data into discriminant function and get the P value of the industry.If it is more than 0.4,then the industry average risk of default probability is bigger, otherwise, is normal.

Table 11. Verification of the training sample

Observed		Predicted		
		y		Percentage Correct
		.00	1.00	
y	.00	72	6	92.3
	1.00	21	2	8.7
Overall Percentage				73.3

Table 11 shows that the average accuracy of industry risk measured by the Logit model is 73.3%, among them the discriminant accuracy of low risk industry is as high as 92.3%, but the discriminant accuracy of high risk industry is only 8.7%.

Table 12. Verification of the predicting sample

Observed		Predicted		
		y		Percentage Correct
		.00	1.00	
y	.00	72	8	92.3
	1.00	19	2	9.5
Overall Percentage				73.3

Taking 101 industry of year 2013 as the predicting sample, the results are shown in table 12.The the average accuracy of industry risk measured by the Logit model is 73.3%.Among them the discriminant accuracy of low risk industry is as high as

90.5%, but the discriminant accuracy of high risk industry is only 9.5%.So the model of low risk industry for determining accuracy is much higher than the determination accuracy of high-risk industry,and the first kind of misjudgment rate was 90%, the second kind of misjudgment rate was 10%.This Logit model to determine results of the training sample and predicting sample almost have the same average rate, so the model prediction effect is good, it is acceptable.

Finally on the basis of the test, using the function (5) to calculate the average risk of default as a result, the result is shown in the table 13.

Table 13. Industry credit risk index (part)

The name of the industry	The credit risk index	The name of the industry	The credit risk index
Salt industry	0.12176	Transportation equipment manufacturing industry	0.19227
Animal husbandry	0.27423	Structural fabricated metal products	0.27003
The ship manufacturing	0.32937	Metal tool manufacturing	0.25568
Electrical equipment manufacturing	0.39282	Metal processing machinery manufacturing	0.2736
Electric manufacturing	0.22309	Manufacture of Metal Products	0.31697
Electric power industry	0.16955	Wine and beverage manufacturing	0.16496
The power supply industry	0.03466	Cigarette Manufacturing Industry	0.27739
Electric power industry	0.04971	Forestry	0.1542
Electrical machinery and equipment manufacturing	0.28995	Bast Textile Industry	0.28675
Telecommunications industry	0.00472	Wool textile industry	0.31791
The electronics industry	0.16955	Coal industry	0.05868
The electronic computer manufacturing	0.14354	Cotton, chemical fiber textile industry	0.28772
Electronic component manufacturing industry	0.27915	Motorcycle manufacturing	0.28369
Textile industry	0.29722	Agriculture, forestry and fishing water conservation machinery manufacturing	0.24588
Fertilizer manufacturing	0.14277	Agriculture,forestry,animal husbandry and fishery industry	0.28208
Clothing, shoes, hats manufacturing	0.3155	pesticide industry	0.2183
Arts and crafts and other products	0.20173	Agriculture	0.20996
Radio and television equipment manufacturing	0.29551	Leather fur down and its products	0.54943
Boiler and prime mover manufacturing	0.28932	Beer manufacturing	0.16209
Ferrous Metals Mining and Dressing	0.20998	Flat glass products	0.19225

These risk indexes of all industries reflect the size of the relative risk, risk results can be interpreted as: the higher credit risk index, the lower the relative concentration limit loan.Results show that the leather fur down and its products, food processing industries are relatively high risk of default, the concentration limit of loan is relatively low.And electric power industry, power industry, telecommunications, petroleum and petrochemical industry and other industries of default risk index is relatively low,its loan concentration limit can be relatively increased.This result can be adjusted for the optimization of the structure of the bank's loan, can also help reduce the risk concentration.

5. Conclusion

This paper builds the index system including the operation and financial risk, industry risk characteristics etc 16 variables indexes and measures the credit risk of Electric power industry, telecommunications, agriculture, fishing, oil processing and coking industry etc 101 industries.This paper takes non-performing asset ratio as the boundary between the high risk and low risk industries.According to the industry statistics from IFinD database,and take the non-performing asset ratio data as the foundation, samples are divided into two groups, 2012 samples are the training sample,2013 samples are the predicting sample.PCA(principal component analysis) method is used for index variables dimension reduction and get three principal components X1,X2,X3,which represents the industry's development capacity, industry's development of the cyclical and the profitability of the industry.Then build Logit model with the X1, X2, X3 as independent variables and the credit risk of industry as dependent variable.Finally do validity test for the Logit model using sample data.The test result is the model of low risk industry for determining accuracy is much higher than the determination accuracy of high-risk industry.And the credit risk results of 101 industries calculated by the Logit model show that leather fur down and its products, food processing industries are higher risk of default;Electric power industry, power industry, telecommunications, petroleum and petrochemical industry and other industries are lower risk of default.

References

[1] S. Michael: *What Can We Learn from Credit Markets.* Proceedings of the 93rd Annual Meeting of the American Law Institute (2016), 1–6.

[2] V. Acharya, I. Drechsler, P. Schnabl, A. P. Victory: *Bank Bailouts and Sovereign Credit Risk.* The Journal of Finance *69* (2014), 2069–2739.

[3] S. Sehgal, T. J. Agrawal: *Bank Risk Factors and Changing Risk Exposures in the Pre- and Post-financial Crisis Periods: An Empirical Study for India.* Management and Labour Studies *42* (2017), No. 4, 356–378.

[4] B. Arindam: *Understanding the Effect of Concentration Risk in the Banks' Credit Portfolio: Indian Cases.* MPRA Paper *9* (2010), No. 1, 1–59.

[5] C. Bluhm, L. Overbeck, C. Wagner: *Introduction to Credit Risk Modeling.* CRC Press (2016), No. 9, 1–59.

[6] T. Beck, O. D. Jonghe, G. Schepens: *Bank competition and stability:cross-country heterogeneity.* Journal of financial intermediation *22* (2013) 218–244.

[7] S. Deng, E. Elyasiani, J. Y. Jia: *Institutional Ownership, Diversification, and Riskiness of Bank Holding Companies.* The Eastern Finance Association *48* (2014), No. 3, 385–415.

[8] I. Edward: *discriminant analysis and the prediction of corporate bankruptcy.* Journal of finance *23* (1968), 589–609.

[9] I. E. Altman, G. Robert, P. Haldeman: *Narayanan,ZETA^{TM} analysis A new model to identify bankruptcy risk of corporations.* Journal of banking and finance *1* (1977), 29–54.

[10] L. Lin, J. Piesse: *Identification of corporate distress in UK industrials:a conditional probability analysis approach.* Applied financial economies *14*, (2004), No. 2, 180–189.

[11] D. Amiram, A. Kalay: *Industry Characteristics, Risk Premiums, and Debt Pricing.* The Accounting Review: January *92* (2017), No. 1, 1–27.

Received November 16, 2017