A Novel DBN-EFA-CFA-Based Dimensional Reduaction for Credit Risk Measurement

ZHANG Yue, HUANG Zhenzhen, SHI Longmei, ZOU Jian
School of Mathematics- Physics and Finance, Anhui Polytechnic University, Wuhu 241000, Anhui, China

Abstract: Affected by the Federal Reserve’s interest rate hike and the downward pressure on the domestic economy, the phenomenon of default is still prominent. The credit risk of the listed companies has become a growing concern of the community. In this paper we present a novel credit risk measurement method based on a dimensional reduaction technique. The method first extracts the risk measure indexes from the basal financial data via dimensional reduaction by using deep belief network (DBN), exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) in turn. And then the credit risk is measured by a systemic structural equation model (SEM) and logistic distribution. To validate the proposed method, we employ the financial data of the listed companies from Q1 2019 to Q2 2022. The empirical results show its effectiveness on statistical evaluation, assessment on testing samples and credit risk forecasting.

Key words: credit risk measurement; dimensional reduaction; deep belief network (DBN); exploratory factor analysis (EFA); confirmatory factor analysis (CFA)

CLC number: O 212.4

0 Introduction

In March 2014, the “11 Chaori Bond” was declared in default due to the inability of ST Chaori to repay its debts and interest, becoming the first bond in China’s history to default on its debts[1]. Since then, defaults in our credit markets have occurred from time to time. In 2020, due to the impact of external environment, the downward pressure of the domestic economy brought huge uncertainty to highly indebted companies, and the scale of credit defaults hit another record high. In 2020, annual bond defaults exceed 50 billion, becoming the highest in nearly five years[2]; in the first half of 2021 alone, the size of the bond is already close to 98.4 billion, substantially exceeding the level of the same period in 2020. On 17 August 2021, the Central Financial and Economic Commission emphasized the need to strengthen the financial rule of law and infrastructure, deepen the construction of credit system and play the fundamental role of credit in the identification, monitoring, management and disposal of financial risks[3]. In such an institutional environment, the study of corporate credit risk is undoubtedly of great relevance and policy implications.

Listed enterprises are the cornerstones of the security market in China. The quality of the enterprises, the
standardization of their conduct and their financial situation will have a direct impact on the development of the Chinese securities market and the interests of investors\cite{7}. As for the research methods of credit risk of listed companies, there are mainly traditional credit risk analysis, multivariate statistical models, and artificial intelligence-based credit risk assessment models\cite{8}. Traditional credit risk analysis research mainly includes SC element analysis method, LAPP (Liquidity, Activity, Profitability, Potentialities) method, and five-level classification method. This type of research method is primarily based on the experience and subjective analysis of bank experts to assess credit risk. In practice, due to the strong subjectivity, the determination of debtor default rates, default factors and their weights relies heavily on the creditors' own judgment, thus making the assessment results of credit risk lack objectivity and scientificity\cite{7}.

Following the traditional methods of credit risk analysis, multivariate statistical models have been widely used abroad in the assessment of credit risk, summarized as linear probability models, logistic models, probit models and discriminant analysis models\cite{8}. With the advent of the Big Data era, neural network technology took a place by virtue of its non-normal distribution and non-linear characteristics, and was introduced into corporate credit evaluation in the 1990s, breaking the inherent limitations of traditional credit risk assessment models, further promoting the development of credit risk assessment models\cite{9}.

The greatest advantage of multivariate statistical models is their obvious explanatory nature, while the strength of neural network techniques lies in the complex combination of a large number on non-linear network layers which allow the extraction of features at various levels of abstraction from the raw data\cite{10}. Therefore, this paper combines deep belief networks, exploratory factor analysis, confirmatory factor analysis, structural equation model with logistic distribution to measure credit risk. With the help of deep belief networks for first-level extraction, the indicators that have a significant impact on credit risk are extracted from the financial indicators, then the indicators are extracted twice by exploratory factor analysis, and thereafter the final extraction of risk measure indexes is achieved by confirmatory factor analysis; after that the risk measure indexes, structural equation model and logistic distribution are used to quantify risk measurement. From there, the potential variables affecting credit risk and the paths of influence behind them (the quantitative structural relationship between each latent variable and credit risk) are identified and a credit risk measurement model is obtained.

1. Establishment of Credit Risk Assessment Method

In the study, the data we obtained may have information redundancy and multicollinearity, so we need to extract the indexes that have a significant impact on the credit risk from the original financial indicators. We obtain the deep nonlinear structure of the input data by deep belief network (DBN) to extract the essential features implied by the input data and perform first-level extraction of the original financial indicators\cite{11}. Then, we analyze how many common factors affect the metrics and the nature of each factor by exploratory factor analysis (EFA), extracting the metrics in the second level after the first level extraction\cite{12}. Finally, the data were analyzed by confirmatory factor analysis (CFA) for structural validity, convergent validity, and discriminant validity to finally extract the metrics we need, i.e., risk measure indexes. Systemic structural equation model (SEM) and logistic distribution are further used to find the specific paths that risk measure indexes affect credit risk and the variables hidden behind these indexes to quantificat credit risk measurement, as shown in the flow chart in Fig. 1\cite{13}.

1.1 Risk Measure Indexes Extraction

Let the financial indicators be $X_1, X_2, \ldots, X_n$. The original risk level is divided into two categories of ST (marked with "0") and non-ST (marked with "1").

In the first step, the risk measure indexes are extracted at the first level by DBN. According to the size of the weight parameters of the neurons in each layer of the network, the $n$ input neurons $X_{i1}, X_{i2}, \ldots, X_{in}$ with the largest weights are found in reverse from the output layer to obtain the first level extraction of the risk measure indexes.

In the second step, the risk measure indexes are extracted at the second level by EFA. For the risk measure indexes extracted at the first level, the metrics with factor loadings greater than 0.7 and commonality greater than 0.4 are extracted according to the factor loadings and commonality of each metric in EFA to obtain the risk measure indexes $X_{i1}, X_{i2}, \ldots, X_{in}$ extracted at the second level.
where $\eta$ is the regression coefficients and $\epsilon_i$ are the error terms.

In the third step, the final extraction of risk measure indexes is performed through CFA. For the risk measure indexes extracted at the second level, it is necessary to pass the $p$-value and standard loading coefficient tests in CFA. At the same time, it is necessary to meet the conditions of average variance extracted (AVE) $>0.5$, composite reliability (CR) $>0.7$, and the AVE square root value is greater than the maximum value of the absolute value of the correlation coefficient between the factors. We will use the indexes that meet all the above conditions as the final obtained risk measure indexes $X_{1i}, X_{2i}, \ldots, X_{ni}$.

### 1.2 Credit Risk Quantification

The common factors obtained from the CFA in the process of risk measure indexes extraction are set as $\xi_1, \xi_2, \ldots, \xi_n$ as latent variables to measure the credit risk $\eta$ of listed companies, their corresponding risk measure indexes are set as $X_{1i}, X_{2i}, \ldots, X_{ni}$ as observed variables to measure the credit risk $\eta$ of listed companies, and the credit risk measurement $X'$ as the observed variable corresponding to the credit risk $\eta$.

Based on the structural equation model, the structural model and measurement model of credit risk are:

\begin{equation}
\eta = l_1 \xi_1 + l_2 \xi_2 + \cdots + l_n \xi_n + \epsilon_i
\end{equation}

\begin{equation}
X' = b \eta
\end{equation}

where $l_1, l_2, \ldots, l_n, b$ are the regression coefficients and $\epsilon_i$ is the error term.

And the measurement model for each obvious and latent variable is:

\begin{equation}
X_{1i} = a_{11} \xi_1 + a_{12} \xi_2 + \cdots + a_{1n} \xi_n + \epsilon_i
\end{equation}

\begin{equation}
X_{2i} = a_{21} \xi_1 + a_{22} \xi_2 + \cdots + a_{2n} \xi_n + \epsilon_2
\end{equation}

\[ \vdots \]

\begin{equation}
X_{ni} = a_{ni} \xi_1 + a_{n2} \xi_2 + \cdots + a_{nn} \xi_n + \epsilon_n
\end{equation}

where $a_{ij}, \ldots, a_{in}$ are the regression coefficients and $\epsilon_i, \epsilon_2, \ldots, \epsilon_n$ are the error terms.

We can obtain:

\begin{equation}
\begin{bmatrix}
\xi_1 = b_{11} X_{1i} + b_{12} X_{2i} + \cdots + b_{1n} X_{ni} + \epsilon_1 \\
\xi_2 = b_{21} X_{1i} + b_{22} X_{2i} + \cdots + b_{2n} X_{ni} + \epsilon_2 \\
\vdots \\
\xi_n = b_{n1} X_{1i} + b_{n2} X_{2i} + \cdots + b_{nn} X_{ni} + \epsilon_n
\end{bmatrix}
\end{equation}

where $b_{ii}, \ldots, b_{nn}$ are the regression coefficients and $\epsilon_i, \epsilon_2, \ldots, \epsilon_n$ are the error terms.

Combining equations (1), (2), (4) yields our risk measurement, i.e.

\begin{equation}
X' = b \eta = b[l_1 (b_{11} X_{1i} + \cdots + b_{1n} X_{ni} + \epsilon_i) + \cdots + l_n (b_{n1} X_{1i} + \cdots + b_{nn} X_{ni} + \epsilon_n)]
\end{equation}

\begin{equation}
= b[l_1 (b_{11} + \cdots + l_n b_{n1}) X_{1i} + \cdots + l_n (b_{n1} + \cdots + l_n b_{nn}) X_{ni} + (l_1 \epsilon_i + \cdots + l_n \epsilon_n)]
\end{equation}

\begin{equation}
= bc_1 X_{1i} + bc_2 X_{2i} + \cdots + bc_n X_{ni} + \epsilon
\end{equation}

where $c_1, c_2, \ldots, c_n$ are the regression coefficients and $\epsilon$ is the error term.

Finally, the credit risk level $X$ is obtained by logistic distribution.

\begin{equation}
p(X = 1 | X') = \frac{1}{1 + e^{-x'}}
\end{equation}

\begin{equation}
p(X = 0 | X') = 1 - p(X = 1 | X') = e^{-x'}/(1 + e^{-x'})
\end{equation}

where $X$ is divided into two classification results ST (marked with "0") and non-ST (marked with "1").

### 2 Experiments

#### 2.1 Data and Data Pre-Processing

Considering data availability and sample representativeness, this paper selects all non-ST companies in the CSI 300 (300 companies) and all ST companies in listed companies (190 companies), totaling 490 companies, as research subjects. The data set is derived from all financial information of these 490 companies in the Wind database for the period Q1 2019 to Q1 2021, totaling 200

![Credit Risk Measurement Model Construction](image)
financial metrics. Since each company had missing data, including 21 financial indicators with all missing data in individual quarters and 207 companies with some financial indicators completely missing, we interpolated by the third spline interpolation method and the K-nearest neighbor interpolation method based on generalized cosine.

2.2 Credit Risk Measure Indexes Extraction

2.2.1 First-level extraction

In our DBN, we select 55% of the data as the training samples and 45% of the data as the testing samples, with 200 financial indicators as the visual layer neurons, and the original risk level is divided into two categories of ST and non-ST. We connect the last hidden layer to the BP neural network and use the two classification results ST (marked with "0") and non-ST (marked with "1") as the output layer neurons. Then we have to determine the quantity of the network's hidden layers and the nodes of each hidden layer by trial-and-error, together with the accuracy of the testing sample to judge the structural generalization performance, and the results are shown in Fig. 2.

![Fig. 2 The accuracy of the testing sample](image)

In setting the nodes of the latest hidden layer, we fix the quantity of neurons in the latest hidden layer to 12, taking into account that the risk assessment level I index is 12. From 200 neurons to 12 neurons, we set 150, 100, 50 as the number of nodes in the previous hidden layer for a wide range of search in order to reduce the loss of information transmitted in the middle. The experiment found a high accuracy rate around 100 and a low accuracy rate below 50 or above 150. So we set the top hidden layer’s node number to 120, 80, 60, 40, and find the best one for the foremost hidden layer from them. And so on, find the best count of nodes of the topmost hidden layer among the three hidden layers and four hidden layers. The results in Fig. 2 show that the recognition performance of the network containing the three-layer hidden layer structure is higher than that of the two-layer hidden layer network structure and the four-layer hidden layer network structure. Because the comparison in the network structure containing two hidden layers found that the best network recognition performance under the structure of the first hidden layer with the nodes number of 60 and 80, therefore, in the network structure containing three hidden layers, a hidden layer was added as the first layer based on the nodes of the second and third hidden layers at 60 and 12, and the first hidden layer nodes were set at 140, 120, 100 and 80 respectively. The experimental results show that the performance of the network is optimal for a structure with 120 nodes in the first hidden layer. Therefore, we finally chose a DBN containing three hidden layers with 120, 60 and 12 nodes for the first-level extraction of the risk measure indexes.

In the first-level extraction of risk measure indexes, we use a top-down approach, starting from the output layer to find the key neurons in the third hidden layer, seeking the six neurons in the third hidden layer that have a significant impact (the highest weight) on each neuron in the output layer, and locating the eight neu-
rons with the highest frequency from them, finally looking in turn to the key neurons in the input layer, i.e. the significant metrics that affect credit risk. Finally we find its corresponding indicators in the input layer, take the 20 indicators with the highest weights, and the results are shown in Table 1.

Table 1  Ranking of indicators

<table>
<thead>
<tr>
<th>Names of index</th>
<th>Label</th>
<th>Index weight</th>
<th>Names of index</th>
<th>Label</th>
<th>Index weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debt-to-long capital ratio</td>
<td>$X_1$</td>
<td>60.68</td>
<td>Return on equity</td>
<td>$X_{11}$</td>
<td>13.98</td>
</tr>
<tr>
<td>Non-current liabilities/total liabilities</td>
<td>$X_2$</td>
<td>26.50</td>
<td>Return on equity weighted</td>
<td>$X_{12}$</td>
<td>13.75</td>
</tr>
<tr>
<td>Net cash from operating activities/current liability</td>
<td>$X_3$</td>
<td>23.36</td>
<td>Retained earnings per share</td>
<td>$X_{13}$</td>
<td>13.47</td>
</tr>
<tr>
<td>Net profit cash cover</td>
<td>$X_4$</td>
<td>20.39</td>
<td>Current assets turnover ratio</td>
<td>$X_{14}$</td>
<td>13.14</td>
</tr>
<tr>
<td>Equity multiplier</td>
<td>$X_5$</td>
<td>19.00</td>
<td>Shareholder free cash flow per share</td>
<td>$X_{15}$</td>
<td>12.44</td>
</tr>
<tr>
<td>Return on assets</td>
<td>$X_6$</td>
<td>16.24</td>
<td>Non-financing net cash flow/total liabilities</td>
<td>$X_{16}$</td>
<td>11.61</td>
</tr>
<tr>
<td>Annualized return on assets</td>
<td>$X_7$</td>
<td>16.16</td>
<td>Cash to meet invest needs</td>
<td>$X_{17}$</td>
<td>11.26</td>
</tr>
<tr>
<td>Interest cover</td>
<td>$X_8$</td>
<td>16.10</td>
<td>EBITDA/gross revenue</td>
<td>$X_{18}$</td>
<td>10.70</td>
</tr>
<tr>
<td>Corporate free cash flow per share</td>
<td>$X_9$</td>
<td>15.12</td>
<td>Net asset value per share</td>
<td>$X_{19}$</td>
<td>9.89</td>
</tr>
<tr>
<td>Annualized return on total assets</td>
<td>$X_{10}$</td>
<td>14.95</td>
<td>Return on total assets</td>
<td>$X_{20}$</td>
<td>8.83</td>
</tr>
</tbody>
</table>

As shown in Table 1, the weight of debt-to-long capital ratio ($X_1$) is 60.68, which is much higher than the weights of other indicators. This is because the debt-to-long capital ratio reflects the long-term capital structure of an enterprise. The smaller the indicator, the lower the degree of debt capitalization of an enterprise, the lower the pressure of long-term debt repayment, the lower the risk of default, and the relatively smaller its credit risk. We can also see that these 20 indicators are more comprehensive in terms of profitability, solvency, operating capacity and development capacity of the company.

2.2.2 Second-level extraction

Deep belief networks narrow the scope for extracting risk measure indexes. Before formally constructing a risk measurement model, we need to find the true risk measure indexes using EFA and CFA.

In the EFA, by first extracting indicators with factor loadings greater than 0.7, removing indicators which has the similar loading on different factors or falls on a specific factor but have factor loadings less than 0.7, and deleting indicators with communality less than 0.4, we obtained 11 indicators, as shown in Table 2.

Then the 11 obtained indicators are again subjected to EFA for finding the latent variables corresponding to each indicator, and the secondary extraction of risk measure indexes is performed.

Table 2  Factor loadings and communality of indicators

<table>
<thead>
<tr>
<th>Names of index</th>
<th>Factor loading</th>
<th>Communalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.860</td>
<td>0.841</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.967</td>
<td>0.968</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.953</td>
<td>0.909</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.979</td>
<td>0.959</td>
</tr>
<tr>
<td>$X_5$</td>
<td>-0.732</td>
<td>0.539</td>
</tr>
<tr>
<td>$X_6$</td>
<td>0.931</td>
<td>0.879</td>
</tr>
<tr>
<td>$X_7$</td>
<td>0.793</td>
<td>0.629</td>
</tr>
<tr>
<td>$X_8$</td>
<td>0.737</td>
<td>0.546</td>
</tr>
<tr>
<td>$X_9$</td>
<td>-0.789</td>
<td>0.626</td>
</tr>
<tr>
<td>$X_{10}$</td>
<td>0.961</td>
<td>0.966</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Names of index</th>
<th>Factor loading</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{11}$</td>
<td>0.984</td>
<td>0.970</td>
</tr>
<tr>
<td>$X_{12}$</td>
<td>0.167</td>
<td>0.030</td>
</tr>
<tr>
<td>$X_{13}$</td>
<td>0.492</td>
<td>0.358</td>
</tr>
<tr>
<td>$X_{14}$</td>
<td>0.114</td>
<td>0.033</td>
</tr>
<tr>
<td>$X_{15}$</td>
<td>0.499</td>
<td>0.262</td>
</tr>
<tr>
<td>$X_{16}$</td>
<td>0.543</td>
<td>0.354</td>
</tr>
<tr>
<td>$X_{17}$</td>
<td>0.664</td>
<td>0.580</td>
</tr>
<tr>
<td>$X_{18}$</td>
<td>0.495</td>
<td>0.406</td>
</tr>
<tr>
<td>$X_{19}$</td>
<td>0.455</td>
<td>0.318</td>
</tr>
<tr>
<td>$X_{20}$</td>
<td>0.637</td>
<td>0.546</td>
</tr>
</tbody>
</table>
From Table 3, we can see that the degree of freedom (df) is 55, Kaiser-Meyer-Olkin measure of sampling adequacy KMO = 0.667 > 0.6 and the observed value of Bartlett’s sphericity test statistic is 33 974.601 with a significance (Sig.) of 0.000 < 0.001, which indicates that these 11 indicators are highly correlated and also indicates that there is some overlap in the credit risk information of listed companies reflected among the 11 indicators, so we need to use EFA to find the potential variables behind these indicators.

To make the extracted latent variables, i.e., factors, more realistic, according to the extraction principle of feature value > 1, the 11 indicators were dimensionalized, meanwhile, the feature value and cumulative contribution rate were calculated. The results are shown in Table 4: there are 5 factors conforming to the extraction principle, and the cumulative contribution of these 5 factors is 80.28%, that is, the variance explained by these 5 factors accounts for 80.28% of the total variance, there is not much information lost by using these 5 factors to reflect the credit risk of listed companies, so these 5 factors can reflect the credit risk of listed companies comprehensively.

### Table 3 KMO and Bartlett’s sphericity test

<table>
<thead>
<tr>
<th>KMO measure of sampling adequacy</th>
<th>Bartlett’s test of sphericity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approx. chi-square</td>
<td>df</td>
</tr>
<tr>
<td>0.667</td>
<td>33 974.601</td>
</tr>
</tbody>
</table>

### Table 4 Total variance explained

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial eigenvalues</th>
<th>Extraction sums of squared loading</th>
<th>Rotation sums of squared loading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Variance contribution rate %</td>
<td>Cumulative variance contribution rate %</td>
</tr>
<tr>
<td>2</td>
<td>2.509</td>
<td>22.808</td>
<td>49.287</td>
</tr>
<tr>
<td>3</td>
<td>1.249</td>
<td>11.359</td>
<td>60.646</td>
</tr>
<tr>
<td>4</td>
<td>1.082</td>
<td>9.835</td>
<td>70.480</td>
</tr>
<tr>
<td>5</td>
<td>1.078</td>
<td>9.800</td>
<td>80.280</td>
</tr>
<tr>
<td>6</td>
<td>0.921</td>
<td>8.371</td>
<td>88.651</td>
</tr>
<tr>
<td>7</td>
<td>0.747</td>
<td>6.788</td>
<td>95.440</td>
</tr>
<tr>
<td>8</td>
<td>0.280</td>
<td>2.548</td>
<td>97.987</td>
</tr>
<tr>
<td>9</td>
<td>0.134</td>
<td>1.222</td>
<td>99.209</td>
</tr>
<tr>
<td>10</td>
<td>0.060</td>
<td>0.546</td>
<td>99.755</td>
</tr>
<tr>
<td>11</td>
<td>0.027</td>
<td>0.245</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Classifying the 11 indicators according to higher loadings, it can be seen from Table 5 that the factor score formula for the first principal factor is:

\[ F_1 = 0.984X_{15} + 0.979X_4 + 0.953X_2 + 0.006X_1 + 0.020X_7 + 0.051X_1 + 0.021X_6 - 0.012X_12 - 0.043X_7 + 0.006X_13 + 0.003X_8 \]

For \( F_1 \) has large loadings of 0.953, 0.979 and 0.984 on three indicators \( X_6, X_1, \) and \( X_{15} \) respectively. As all three indicators reflect the profitability of listed companies, \( F_1 \) is defined as profitability in this paper.

The factor score formula for the second principal factor is:

\[ F_2 = 0.013X_{20} + 0.024X_7 - 0.005X_3 + 0.967X_4 + 0.961X_7 + 0.097X_3 + 0.318X_1 - 0.009X_{12} - 0.023X_{15} - 0.024X_6 \]

For \( F_2 \) is loaded on 2 items, \( X_1 \) and \( X_{17} \), with loading values of 0.96 or more. Both indicators reflect the cash flow of listed companies, so \( F_2 \) is defined as cash flow.

The factor score formula for the third principal factor is:
Rotated component matrix

<table>
<thead>
<tr>
<th>Names of index</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$F_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{20}$</td>
<td>0.984</td>
<td>0.013</td>
<td>0.024</td>
<td>0.019</td>
<td>0.002</td>
</tr>
<tr>
<td>$X_7$</td>
<td>0.979</td>
<td>0.024</td>
<td>0.031</td>
<td>0.017</td>
<td>0.002</td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.953</td>
<td>-0.005</td>
<td>0.030</td>
<td>0.010</td>
<td>0.001</td>
</tr>
<tr>
<td>$X_6$</td>
<td>0.006</td>
<td>0.967</td>
<td>0.180</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$X_{17}$</td>
<td>0.020</td>
<td>0.961</td>
<td>0.205</td>
<td>0.012</td>
<td>0.002</td>
</tr>
<tr>
<td>$X_5$</td>
<td>0.051</td>
<td>0.097</td>
<td>0.931</td>
<td>-0.005</td>
<td>0.017</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.021</td>
<td>0.318</td>
<td>0.860</td>
<td>0.012</td>
<td>-0.023</td>
</tr>
<tr>
<td>$X_{13}$</td>
<td>-0.012</td>
<td>0.000</td>
<td>0.028</td>
<td>0.793</td>
<td>0.006</td>
</tr>
<tr>
<td>$X_{16}$</td>
<td>-0.043</td>
<td>-0.009</td>
<td>0.024</td>
<td>-0.789</td>
<td>0.004</td>
</tr>
<tr>
<td>$X_{15}$</td>
<td>0.006</td>
<td>-0.023</td>
<td>0.049</td>
<td>0.007</td>
<td>0.737</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.003</td>
<td>-0.024</td>
<td>0.052</td>
<td>0.005</td>
<td>-0.732</td>
</tr>
</tbody>
</table>

$F_1 = 0.024X_{20} + 0.031X_7 + 0.030X_6 + 0.180X_4$ 
$+0.205X_{17} + 0.931X_5 + 0.860X_2 + 0.028X_{12}$ 
$+0.024X_{16} + 0.049X_{15} + 0.052X_8$

The loadings of $F_1$ in $X_7$ and $X_6$ are larger, one exceeds 0.9 and one is above 0.85. These 2 indicators can reflect the solvency of listed companies, so $F_1$ is defined as solvency.

The factor score formula for the fourth principal factor is:

$F_4 = 0.019X_{20} + 0.017X_7 + 0.010X_6 + 0.012X_{17}$ 
$-0.005X_5 + 0.012X_4 + 0.793X_{12}$ 
$-0.789X_{16} + 0.007X_{15} + 0.005X_8$

$F_4$ loads more on $X_{12}$ and $X_{16}$, the former is 0.793, reflecting the profitability and operation management level of listed companies, the latter is -0.789, reflecting the short-term solvency of listed companies and reflecting the goal of listed companies’ operation pursuit, so here $F_4$ is defined as the operation level.

The factor score formula for the fifth principal factor is:

$F_5 = 0.002X_{20} + 0.002X_7 + 0.001X_6 + 0.002X_{17}$ 
$+0.017X_5 - 0.023X_4 + 0.006X_{12}$ 
$+0.004X_{16} + 0.737X_{15} - 0.732X_8$

The loadings of $F_5$ on $X_4$ and $X_6$ are large, and their absolute values are above 0.7. These two indicators can reflect the development ability of listed companies, so $F_5$ is defined as development ability. According to the results of the aforementioned factor analysis, the potential variables corresponding to these 11 indicators are these 5 main factors, and the final extraction of the risk measure indexes can be discussed through these 5 main factors.

2.2.3 Final extraction

According to Occam’s Razor, the "simple and effective principle", our model should be as simple as possible. In order to find indicators that can portray the degree of risk more quickly and concisely, we performed validity analysis on the data through CFA on the basis of EFA, and then further removed the inappropriate variables to achieve the final extraction of risk measure indexes.

It can be seen from Table 6 that when $X_{12}$ measures the operation level, the absolute value of its standardized load coefficient is 0.254<0.6, which means that the measurement relationship is weak. When $X_{16}$ measures the operation level, the standardized load coefficient does not show significant ($p=0.378>0.05$), indicating that there is no significant measurement relationship. The absolute values of the standardized load coefficients of $X_4$ and $X_{15}$ are 0.52 and 0.151 respectively, both less than 0.6, indicating that the measurement relationship is weak.

AVE and CR are used for convergent validity analysis. As can be seen from Table 7, the CR corresponding to operation level is less than 0.7. The AVE corresponding to development ability is less than 0.5, and the CR is less than 0.7. These imply that the aggregated validity of the data for this analysis is poor, and that the operation level and development capacity needs to be removed for further analysis. Combining Table 6 and Table 7, it can be seen that the observed variables corresponding to the operation level and development capacity need to be deleted.

After deleting $X_{12}, X_{16}, X_{15}$, and $X_{16}$, we performed CFA again on the remaining indicators, and the results were as follows. For the measurement relationship, the absolute values of the standardized load coefficients were all greater than 0.6 and appeared significant, implying a good measurement relationship. In terms of AVE and CR, all three factors corresponded to AVE values greater than 0.5, and all CR values were higher than 0.7, implying that the data of this analysis had good convergent validity. From the analysis of Table 8 for discriminant validity, for profitability, the AVE square root value is 0.96, which is greater than the maximum value of the absolute value of the inter-factor correlation coefficient of 0.07, implying a good discriminant validity. Similarly, it can be seen that solvency and cash flow also have good discriminant validity.
A VE and CR

through its impact on profitability; itability indirectly affects the credit risk of listed companies and each indicator of profitability is the risk measurement). This is as observed variables. The observed variable corresponding to credit risk is the risk measurement ($X'\)'. This is assumed as follows:

**H1**: Profitability has a significant impact on the credit risk of listed companies and each indicator of profitability indirectly affects the credit risk of listed companies through its impact on profitability;

**H2**: Cash flow has a significant impact on the credit risk of listed companies and each indicator of cash flow indirectly affects the credit risk of listed companies through its impact on cash flow;

**H3**: Solvency has a significant impact on the credit risk of listed companies, and each indicator of solvency indirectly affects the credit risk of listed companies through its impact on solvency.

### 2.3 Risk Measurement

#### 2.3.1 SEM assumptions

Through the extraction of risk measure indexes, we found that profitability, cash flow and solvency are latent variables to measure the credit risk of listed companies, and their corresponding risk measure indexes are observed variables. The observed variable corresponding to credit risk is the risk measurement ($X'\)'. This is assumed as follows:

**H1**: Profitability has a significant impact on the credit risk of listed companies and each indicator of profitability indirectly affects the credit risk of listed companies through its impact on profitability;

**H2**: Cash flow has a significant impact on the credit risk of listed companies and each indicator of cash flow indirectly affects the credit risk of listed companies through its impact on cash flow;

**H3**: Solvency has a significant impact on the credit risk of listed companies, and each indicator of solvency indirectly affects the credit risk of listed companies through its impact on solvency.

#### 2.3.2 Credit risk path analysis and risk measurement

In order to further verify the rationality of the model and the impact of each indicator, the SEM of credit risk for listed companies is constructed based on the above assumptions, and its impact path is shown in Fig. 3.

In Fig. 3, we can see that the exogenous variables include profitability, cash flow, and solvency, which are set as $\xi, \xi_2, \xi_3$ . The endogenous variable is credit risk, set as $\eta$. Based on the SEM principle, the mathematical expression of the theoretical model can be obtained.

The structural and measurement models for credit risk are:

$$\eta = 0.39\xi + 0.061\xi_2 + 0.094\xi_3$$

$$X' = 0.642\eta$$

As can be seen from the model, profitability, solvency and cash flow all have different degrees of influence on the credit risk of listed companies. Among them, profitability has the most significant impact with a regression coefficient of 0.39, while solvency and cash flow have the second most significant impact with regression coefficients of 0.094 and 0.061, respectively.
Profitability is measured by the following model:

\[ X_7 = 0.978 \xi_1, \quad X_6 = 0.905 \xi_1, \quad X_3 = 0.994 \xi_1 \]

The metric model for profitability is:

\[ \xi_1 = \frac{1}{0.978 + 0.905 + 0.994} X_7 + \frac{1}{0.978 + 0.905 + 0.994} X_6 + \frac{1}{0.978 + 0.905 + 0.994} X_3 \]

Among the profitability, the most significant impact is the return on total assets with a path coefficient of 0.994. This indicates that the return on total assets is the key to corporate profitability. Combined with the structural model of credit risk it is clear that hypothesis H1 holds, i.e. profitability has a significant impact on the credit risk of listed companies and that the indicators of profitability indirectly affect the credit risk of listed companies through their impact on profitability.

Cash flow is measured by the following model:

\[ X_4 = 0.952 \xi_2, \quad X_{17} = 0.987 \xi_2 \]

The metric model for cash flow is:

\[ \xi_2 = \frac{1}{0.952 + 0.987} X_4 + \frac{1}{0.952 + 0.987} X_{17} \]

From the perspective of cash flow, the path coefficient of cash to meet invest needs is larger, which is 0.987. The larger the indicator is, the higher the self-sufficiency rate of the enterprise’s capital, and the stronger the enterprise’s ability to maintain production and operation. Combined with the structural model of credit risk it is easy to know that hypothesis H2 holds, that is, cash flow has a significant impact on the credit risk of listed companies, and each indicator of cash flow indirectly affects the credit risk of listed companies through the impact on cash flow.

Solvency is measured by the following model:

\[ X_9 = 0.697 \xi_3, \quad X_1 = \xi_3 \]

The metric model for solvency is:

\[ \xi_3 = \frac{1}{0.697 + 1} X_9 + \frac{1}{0.697 + 1} X_1 \]

In terms of solvency, the path coefficients for corporate free cash flow per share and net cash from operating activities/current liability are 0.697 and 1, respectively.
Net cash from operating activities/current liability is more heavily weighted and the ratio provides a cash flow perspective on a company’s ability to pay off short-term liabilities in the current period. Therefore, hypothesis H3 holds that solvency has a significant impact on the credit risk of listed companies and that the solvency indicators indirectly affect the credit risk of listed companies through their impact on solvency.

In summary, the risk measurement is:
\[ X' = 0.0248X_1 + 0.0199X_2 + 0.0883X_3 + 0.0817X_4 + 0.0356X_5 + 0.0192X_6 + 0.0804X_7 \]
The risk level can be determined:
\[ p(X = 1 | X') = 1/(1 + e^{-X'}) \]
\[ p(X = 0 | X') = 1 - p(X = 1 | X') = e^{-X'}/(1 + e^{-X'}) \]

The standardised coefficient estimates, standard errors, critical ratios and p-values for the main SEM pathways can be seen in Table 9. In the measurement model, it can be seen that the regression coefficients of the observed variables corresponding to each latent variable are all above 0.5 and all are significant at the 0.05 confidence level. In the structural model, the regression coefficients between the latent variables also all passed the significance test at the 0.05 confidence level. This indicates that the model meets the basic fit criteria and is identifiable.

| Table 9  Summary table of model regression coefficients |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
| \( X \rightarrow Y \) | Standard error | Critical ratio | \( p \)-value | Standard load coefficient |
| Profitability → Credit risk  | 0.016           | 17.035          | 0.000          | 0.390            |
| Cash flow → Credit risk      | 0.017           | 2.364           | 0.018          | 0.061            |
| Solvency → Credit risk       | 0.017           | 3.624           | 0.000          | 0.094            |
| Profitability → \( X_1 \)    | 0.008           | 127.620         | 0.000          | 0.978            |
| Profitability → \( X_6 \)    | –               | –               | –              | –                |
| Profitability → \( X_{20} \) | 0.008           | 134.540         | 0.000          | 0.994            |
| Cash flow → \( X_1 \)        | –               | –               | –              | –                |
| Cash flow → \( X_7 \)        | 0.012           | 85.397          | 0.000          | 0.987            |
| Solvency → \( X_6 \)         | 0.023           | 29.933          | 0.000          | 0.697            |
| Solvency → \( X_7 \)         | –               | –               | –              | –                |
| Credit risk → \( X' \)       | –               | –               | –              | 0.642            |

2.4 Performance Evaluation

2.4.1 Statistical evaluation

In this section, to further validate the soundness of the model, we evaluate the structural equation model based on the model fit index and the following table shows the results.

The SEM fit index test results are shown in Table 10, and the fit criteria are basically met, except for \( \chi^2 \) and \( \chi^2/\text{df} \), which are not ideal. However, these two indicators are susceptible to sample size and tend to reject all well-fitting models when the sample size is large, so the model is not revised here.

2.4.2 Assessment on testing samples

In this section we assess the credit risk of the data in the testing samples with a risk metric model. The testing samples accounts for 45% of the original sample, with a total of 1 984 samples and 13 892 data, with 2 039 missing data, and a missing rate of 14.68%. Among them, the data of cash to meet invest needs indicator is missing for 5 quarters, and the data of net cash from operating activities/current liability indicator is missing for 27 companies. We interpolate these missing data by data preprocessing. It can be seen that incomplete data information is destined to lead to a lack in our evaluation, and the results are shown in Table 11.

As can be seen from Table 11, 80 of the non-defaulted samples were correctly assessed, 641 of the defaulted samples were correctly assessed, and the overall prediction accuracy of the risk metric model was \((802 + 641)/1984 = 72.73\%\). In the evaluation, the type I error rate, i.e., the error rate of determining the non-defaulted sample as the defaulted sample, can be found to be \(408/(802+408) = 33.72\%\). The type II error rate, i.e., the error rate of determining the defaulted sample as the non-
A Novel DBN-EFA-CFA-Based Dimensional Reducation for Credit Risk Measurement

Assessment on testing samples

There were a total of 514 samples of predicated. The Type I error rate was 432/(1 075+432) = 28.67% and the Type II error rate was 4.16% lower than the Type I error rate, at 187/(187+576) = 24.51%. There were a total of 1 651 samples of prediction data, of which 619 were predicted correctly. Of the non-defaulted samples, a total of 432 samples were judged incorrectly and judged to be defaulted. Of the defaulted samples, a total of 187 samples were determined to be incorrect and judged to be non-defaulted.

3 Conclusion

This paper uses financial data from Q1 2019 to Q2 2022 as the research data set for construction and evaluation of a novel credit risk measurement model. The measure indexes for credit risk of listed companies are obtained through DBN, EFA and CFA in turn. Then the risk measurement is obtained with SEM and logistic distribution. Finally, the performance evaluation of the risk measurement is made with statistical evaluation, assessment on testing samples and credit risk forecasting, resulting in the following conclusions:

1) The credit risk of listed companies is mainly influenced by profitability, cash flow and solvency. Profitability has the greatest influence with a regression coefficient of 0.39, cash flow has the least influence with a regression coefficient of 0.061, and solvency is in the middle with a regression coefficient of 0.094.

2) Among profitability, the return on assets has the highest impact weight, annualized return on assets as the research data set for construction and evaluation of a novel credit risk measurement model. The measure indexes for credit risk of listed companies are obtained through DBN, EFA and CFA in turn. Then the risk measurement is obtained with SEM and logistic distribution. Finally, the performance evaluation of the risk measurement is made with statistical evaluation, assessment on testing samples and credit risk forecasting, resulting in the following conclusions:

3) Cash flow is mainly reflected by the net profit cash cover and the cash to meet invest needs, which has an impact weighting of 0.509 0 and 0.491 0, respectively.

4) Within solvency, the impact of corporate free
References


