Research Article

Discriminant Analysis of Zero Recovery for China's NPL

Yue Tang, Hao Chen, Bo Wang, Muzi Chen, Min Chen, and Xiaoguang Yang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100090, China

Correspondence should be addressed to Xiaoguang Yang, xgyang@iss.ac.cn

Classification of whether recovery of non-performing loans (NPL) is zero or positive is not only important in management of non-performing loans, but also is essential for estimating recovery rate and implementing the new Basel Capital Accord. Based on the largest database of NPL's recovering information in China, this paper tries to establish discriminant models to predict the loan with zero recovery. We first use Step-wise discrimination method to select variables; then give an in-depth analysis on why the selected variables are important factors influencing whether a loan is zero or positive recovery rate. Using the selected variables, we establish two-type discriminant models to classify the NPLs. Empirical results show that both models achieve high prediction accuracy, and the characteristics of obligors are the most important factors in determining whether a NPL is positively recovered or zero recovered.

1. Introduction

Credit risk is the major risk commercial banks are faced with, hence, measurement and management of credit risk is the core task of risk management. In China, commercial banks' credit risk mainly features the accumulation of a large number of nonperforming loans (NPLs), so it is critical to model various risk factors in NPLs in order to establish a sound credit risk management system.

Loss given default (LGD) is an equivalent concept with recovery rate. LGD is a critical parameter for measurement of credit risk, a basis for estimating expected loss (EL) and unexpected loss (UL). Modeling the LGD of NPLs effectively is very important for the management of NPLs and banking supervision.

The key issues of LGD (Schuermann [1]) are the following: definition and measurement; key drivers, modeling, and estimation approaches. The following studies explored the characteristics of Bond LGD.

Altman and Kishore [2], for the first time, documented the severity of bond defaults stratified by Standard Industrial Classification sector and by debt seniority. They found that the highest average recoveries came from public utilities (70%) and chemical, petroleum, and related products (63%). They concluded that the original rating of a bond issue as investment grade or below investment grade had virtually no effect on recoveries once seniority was accounted for. Neither the size of the issue nor the time to default from its original date of issuance had any association with the recovery rate. Acharya et al. [3] found that recoveries on individual bonds were affected not only by seniority and collateral but also by the industry conditions at the time of default. Altman et al. [4] analyzed the impact of various assumptions about the association between aggregate default probabilities and the loss given default on bank loans and corporate bonds and sought to empirically explain this critical relationship. They specified and empirically tested for a negative relationship between these two key inputs to credit loss estimates and found that the result was indeed significantly negative.

Currently, the bank loan LGD is not explored well by theoretical and empirical literature. In many cases, the subject of many studies on LGD is bond rather than loan. Little research has been done on the LGD of loans. There are almost no public loans' LGD models. Some of the most important research focusing on the bank loan markets is the following.

Carty and Lieberman [5] measured the recovery rate on a sample of 58 bank loans for the period 1989–1996 and reported skewness toward the high end of price scale with the average recovery of 71%. Gupton et al. [6] reported higher recovery rate of 70% for senior secured loans than for unsecured loans (52%) based on 1989–2000 data sample consisting of 181 observations.

The above studies focused on the U.S market. Hurt and Felsovalyi [7] who analyzed 1149 bank loan losses in Latin America over 1970–1996 found the average recovery rate was 68%. None of the above studies provided the information on predicting bank loan LGD.

Moody has established special LGD prediction models, which are known as LossCalc models. But most of the data sample used in the models is bond data; the sample includes only a small portion of the loan data. More details about LossCalc can be found in Gupton [8]. However, the NPLs in China's commercial banks have their own special characteristics. The factors affecting NPL recovery are also not the same with abroad. Therefore, it is essential to develop specific LGD models in accordance with the actual situation of China's commercial bank.

Now, domestic research on LGD is mainly concentrated in the qualitative discussion. Chen [9] discussed the importance of LGD. Moreover, he introduced several international modeling methods for LGD and discussed the major difficulty in modeling LGD in China.

Wu [10] analyzed the necessary conditions for banks to estimate LGD and explored the methods to estimate LGD under the Internal Ratings Based (IRB) framework. Liu and Yang [11] summarized and analyzed the international discussion about the performance of LGD and factors impacting LGD. Moreover, he thought that data was the basis for modeling LGD and suggested that in order to make up for deficiency in data, the central bank should take the lead to establish the joint NPL database. Shen [12] compared many new methods of modeling LGD, including the nonparametric and neural network models.

Studies mentioned above present valuable guidance for research on relevant domestic issues. But these studies are limited to qualitative discussions, so they cannot provide substantive deeper conclusions.

Ye and Liu [13] have carried out some empirical research with the relevant data from banks. With the data on defaulted loans from a bank in Wenzhou, a city in Zhejiang Province,
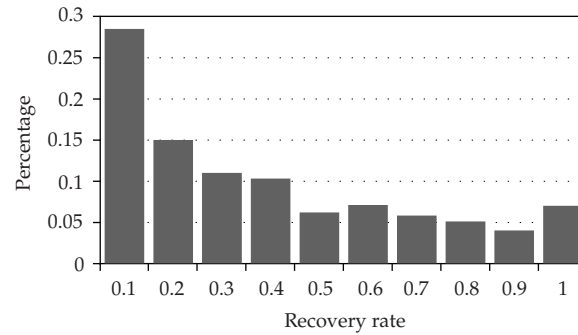
Figure 1: Distribution of recovery rate.

Ye and Liu [13] analyzed the LGD distribution characteristics of NPLs in terms of exposure at default, maturity structure, mortgage manner, industry, and so forth.

Empirical studies mentioned above laid a good foundation for domestic research on LGD, but they are limited by their very small data sample which does not reflect the national situation. Moreover, the analysis tools in the above studies are just simple descriptive statistics. It is hard to convince people from their results.

Data samples used in this paper come from the LossMetrics Database. LossMetrics is a large-scale database concerning NPLs, established by Doho Data Consulting Corporation. This database consists of rich and detailed information about NPLs coming from the four major commercial banks in China, namely, Industrial and Commercial bank of China (ICBC), China Construction Bank (CCB), Agricultural Bank of China (ABC), and Bank of China (BOC). Nearly tens of thousands of obligors' pieces of information are included in this database and the amount of NPLs comes to 600 billion Renminbi(RMB). These NPLs cover more than 20 industries and 10 provinces and can reflect the overall situation and distribution characteristics of China's NPLs.

Research both at home and abroad shows that the recovery rate appears as a bimodal distribution. According to Gupton [8], the recovery rate has peaks in the vicinity of 20% and 80%, respectively. But with China's data analysis, the peaks are located in the vicinity of 0 and 100%, respectively; the peak near zero is very high, while the peak near 100% is relatively low. According to the results of statistical data, the mean of NPLs' recovery rate in China's AMC is about 30%, which is much lower than the recovery rate in developed countries. More than 25% NPLs' recovery rate is 0, and nearly 10% NPLs' recovery rate is almost 100%. Figure 1 shows the distribution of recovery rate of NPLs in China.

Considering the bimodal features of recovery rate's probability distribution, estimating recovery rate directly will lead to a big bias.

It is effective and necessary to model recovery rate in two steps. The first step is to classify recovery into different categories: zero recovery and positive recovery. This paper devotes to the classification. This classification is not only a part of estimating LGD, more importantly if the recovery of an NPL can be determined to be zero or positive, the information can help commercial banks and AMCs to manage and dispose NPLs as they can allocate more resources to assets with positive recovery and avoid wasting money and time on loans with zero recovery, which help reduce financial cost and improve management efficiency.

In this paper, discriminant analysis method is used to classify recovery. Basic idea of discriminant analysis can be illustrated as follows: assuming that the research subject can

be classified into several categories, and there are observation data of given samples in each category. Based on this assumption, we can establish discriminant functions according to some criteria, and use the function to classify the samples with unknown category. There are plenty of discriminant analysis methods, such as Distance, Fisher, Sequential, Bayesian, and Stepwise discrimination. This paper applies Stepwise, Distance, and Bayesian discrimination to classify the recovery of NPLs and compares the results between Distance discrimination and Bayesian discrimination as well. More details about discriminant analysis can be found in Bruch [14].

Rest of the paper is organized as follows. Section 2 describes the data used to model and selects significant variables for classifying recovery by Stepwise discrimination. Section 3 presents an in-depth analysis why the selected variables have a significant influence on discrimination of zero recovery and positive recovery. Section 4 establishes several discrimination models and utilizes these models to predict. Section 5 concludes and discusses pros and cons of the current research and puts forward expectations and extensions for future research.

2. Data and Variable Selection

2.1. Data Samples

Generally speaking, an obligor can borrow one or several loans. Banks or AMCs collect NPL from obligors in a way they collect all the loans from an obligor as a whole. To give a clear classification, we first sift the obligors with only one loan from the database and start our analysis and modeling process with these obligors as samples; then we use the sample of obligors with several loans to test if the models still apply.

From LossMetrics database, we obtain 625 obligors with only one loan, among which, 425 obligors have positive recovery while 200 other obligors have zero recovery. The number of the obligors with several loans is 821, out of which, 592 have positive recovery, and 229 do not have recovery. The total number of loans mounts to be 4061.

2.2. Variable Selection

The variables associated to NPL can be divided into, broadly, two categories: debt information and obligor information.

Debt information refers to the five-category assets classification for bank loans, collateral, NPLs' transfer mechanism, and so forth. The five-category assets classification and for bank loans is a standard established by China's commercial banks, which can be referred to assess the quality of loans. It includes five categories: normal, concerned, subordination, doubt, and loss. Since all the loans studied are nonperforming loan, the samples don't cover the loans belonging to normal or concerned category. NPLs' transfer mechanism refers the ways to transfer NPLs from commercial banks to AMCs. One way is policy-off, that is, the central government buys NPLs from the commercial banks and gives them free to AMCs, and the other way is business acquisition, that is, AMCs buy NPLs directly from commercial banks.

Obligor information refers to obligor's business situation, the industry, and region obligor being in. Business situation can be divided into seven categories: normal, financial distress, ceasing-business, shutting-down, under-insolvency-process, bankruptcy, and unknown.

Table 1: The result of variable selection.

Significant variables	Partial-r square
Loss-category loan	0.4882
Bankruptcy	0.1611
Shutting-down	0.1229
Ceasing-business	0.061
Collateral loan	0.0437
Under-insolvency-process	0.0406
Unknown	0.0241
Business acquisition	0.0142
Tianjin area	0.0148
Fujian area	0.0132
Real estate industry	0.0128
Hebei area	0.0079
Retail industry	0.0054

As there are very few data in the database containing financial indices of the obligors, we can only use the indicator variables in the modeling process.

In order to avoid too many variables to interfere the effect of discrimination, we employ stepwise discrimination method to select the significant variables.

The result of variable selection is displayed in Table 1.

As Table 1 shows, the larger the Partial-r square of a variable is, the more important the variable is. In Table 1, the impact of five-category classification for loans, the business situation of companies, region, and industry is significant. while business situation of company and the quality of loans (five-category classification) play a leading role in discrimination with a proportion up to 94%, in contrast, the factors such as industry and region turn out to be less important.

3. Analysis of Selected Variables

In the above section, we use Stepwise discrimination to select some variables which are significant for distinguishing zero recovery and positive recovery from the view of discriminant function. In order to have a deeper understanding of these variables and hence to be more confident in predicting models in next section, we direct a further probe in this section to show that these variables have bigger impact on zero recovery of NPLs than other unselected variables.

In the linear discriminant function developed by Distance method, the sign of the coefficient of the variable indicates whether the variable acts positively or negatively; however, Bayesian discrimination can compute the probability of the recovery of a loan being zero or positive; therefore, it is more convenient and more accurate to use Bayesian method, since the changes in variables can be connected with the probability. Here, we use Bayesian method to establish linear discrimination function and see how these variables affect zero recovery.

The basic idea of Bayesian discriminant is as follows. Let the Bayesian linear discriminant function for positive recovery be $f_1 = a + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$, and the Bayesian linear discriminant function for zero recovery be $f_2 = b + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$. For

Table 2: Bayesian discriminant function.

Variables	Coefficient 1	Coefficient 12	Difference
Intercept	−2.389	−8.454	6.065
Loss-category loan	1.834	6.851	−5.018
Bankruptcy	1.656	10.521	−8.866
Shutting-down	1.785	7.706	−5.922
Ceasing-business	1.755	5.819	−4.064
Collateral loan	1.743	−0.110	1.853
Under-insolvency-process	0.196	8.673	−8.478
Unknown operation	2.700	5.593	−2.894
Policy-off	3.683	2.093	1.590
Tianjin area	3.339	0.633	2.706
Fujian area	1.694	−1.223	2.917
Real estate industry	1.966	−0.226	2.192
Hebei area	1.617	0.539	1.079
Retail industry	1.901	1.127	0.774

an NPL, the posterior probability of positive recovery will be $p_1 = (e^{f_1}/e^{f_1} + e^{f_2}$ as $e^{f_1})$, and the posterior probability for zero recovery will be $p_2 = (e^{f_1}/e^{f_1} + e^{f_2}$ as $e^{f_2})$. When $p_1 > p_2$, that is $p_1/p_2 = e^{c+c_1x_1+c_2x_2+\cdots+c_nx_n} > 1$, the NPL will be classified into the positive category, or else zero category. Here, $c_k = a_k − b_k$; if $c_k < 0$, this variable will have a positive influence on the NPL and is to be classified into positive category. That is, if this variable changes from 0 to 1 and the other variables keep constant, the probability for positive recovery will become larger; if $c_k > 0$, the probability will become smaller. Moreover, if a variable x_k changes from 0 to 1, the change of the probability for positive recovery will be e^{c_i} times of the probability for zero recovery. Therefore Bayesian linear discriminant function can be used in analyzing the classification efficiency of variables.

Using the data samples obtained from Section 2, we establish the Bayesian discriminant function as follows.

These variables can be classified into 6 classes: five-category classification, collateral properties, transfer mechanism of NPLs, operation situation of the obligors, industry, and region. Now, let us see how these variables affect the classification of NPLs.

3.1. Five-Category Loan Classification

The variable five-category classification plays a leading role in classification based on the results of stepwise discrimination. Only "loss category loan" has entered into the model, because the effects of "subcategory loan" and "doubt-category loan" for classification have been mainly reflected by "loss-category loan".

In the discriminant function, the difference of the coefficient of this variable equals −5.018, which shows that when a nonperforming loan belongs to loss category, the probability of its recovery being zero will be larger than the probability of being positive. Keep the other variables constants; if a loan enter a loss category from other categories, the changes of the probability for zero recovery are $e^{5.018}$ times of the changes of the probability. Combined with Figure 2 it will help understand why five-category classification has played such an
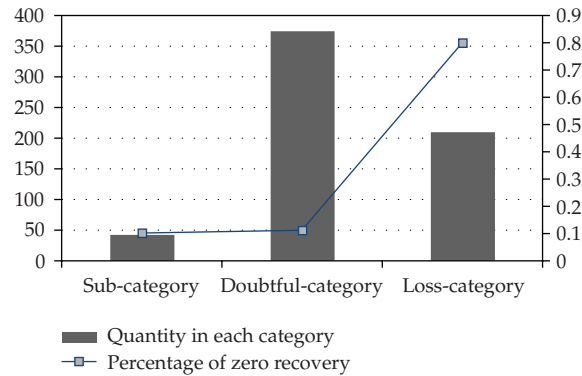
Figure 2

important role. In Figure 2 , the blue point represents the ratio of the NPLs not having recovery in each category.

In Figure 2 , the ratio of zero recovery has decreased along with the grade of loans. The ratio for subcategory and doubt-category is very low, less than 10%, while the ratio for loss-category loans is up to 80%. It shows that, first, whether a loan has recovery is mainly determined by the quality of loan. The worse its quality is, the larger the probability of zero recovery will be. Second, the standard of five-category classification built by China's commercial bank can differentiate the quality of loans effectively. So we should continue to pay much attention to the change of the grade of loans and improve the ability to distinguish the qualities of loans, which will help manage credit risk more effectively.

3.2. Business Situation of Obligors

The significant variables of this type are bankruptcy, under-insolvency-process, shutting-down, and ceasing-business. The variables "normal" and "financial distress" do not enter into the model due to multicollinearity between these variables. And obviously the ability to repay in both cases will be much greater than the other cases.

The differences of the coefficients for these significant variables between the two functions are negative, which means that the recovery rate of NPLs in these cases will be inclined to be classified into zero. Let us see Figure 3.

In Figure 3, we can see that there are significant differences among the ratios of zero recovery in different cases. When the obligors run well or in financial distress, the ratio of their NPLs with positive recovery is very high. Especially when the obligors operate well, the recovery rates of NPLs in this category are all positive.

When the business situation of an obligor is in a state of bankruptcy, under-insolvency-process, shutting-down, or ceasing-business, the ratio of their NPLs with zero recovery is very high; the highest is the obligors in the state of bankruptcy.

We can draw some conclusions from these results: the business situation of obligors reflects their payment ability to a great extent. The worse the business situation is, the weaker the payment ability will be, and vice versa. When obligors enter the states of bankruptcy, shutting-down, or ceasing-business, they have no ability to repay their loans, which will lead to high ratio of NPLs with zero recovery. While obligors run well or even if they are in financial distress, the cash flow of business is still in existence, which makes the probability
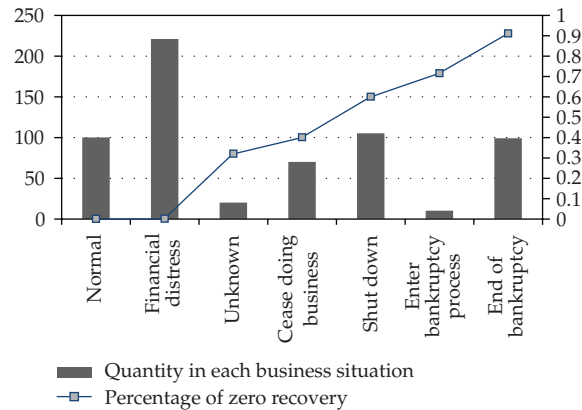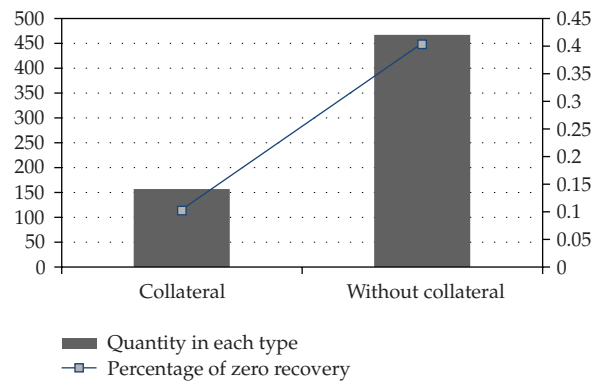
Figure 3

Figure 4

of recovery much larger. So when disposing NPLs, the business situation of enterprises should be paid much attention to. If the business situation can change with time, the NPLs of enterprises should have been disposed when the enterprises run relatively well as far as possible, which has positive effect.

3.3. Collateral

Due to the quality of the data, we cannot get the collateral value; so only the state variable, that is, "whether the loan has effective mortgage", is contained in the model.

In the discriminant function, whether the loan has effective collateral is significant, and the difference between the coefficients in two types of discriminant function is 1.961. This shows that effective collateral has a positive effect in obtaining positive recovery. Keep the other variables constant; when a loan changes from noncollateral to collateral, the change in the probability of positive recovery is $e^{1.961}$ times of the change in the probability of zero recovery. Figure 4 has made this point more clearly in a straight way.

As Figure 4 shows, the ratio of collateral NPLs with zero recovery is 10%; this ratio is approximate 44% for the NPLs without collateral. This result shows that with effective
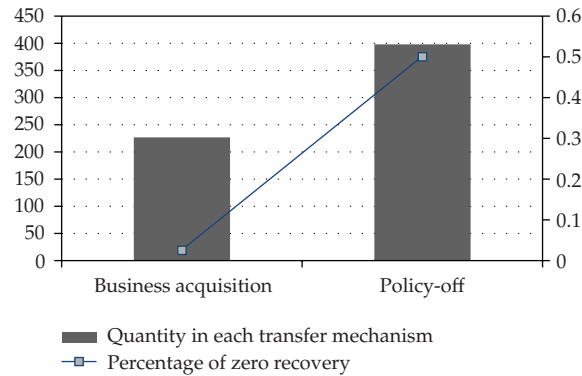
Figure 5

collateral, the probability of positive recovery will increase greatly. This is because the collateral can be used to negotiate with the obligor or sold straightly in the secondary market. When purchasing, managing, and disposing NPLs, the collateral should be paid much attention to, which might determine whether NPLs have recovery and how much could be got from the obligors.

3.4. Transfer Mechanism of NPLs

The variable "policy-off" is significant in Stepwise discrimination. The difference of coefficient between the two discriminant functions is 1.174, which implies that the probability of positive recovery will be larger than that of zero recovery when NPLs are commercially bought from banks. That is, the quality of NPLs by business acquisition may be better than the quality of NPLs by policy-off. More details can be seen in Figure 5.

As is shown in Figure 5, the ratio of NPLs by business acquisition with positive recovery is almost 100%, while the ratio is much lower for NPLs by policy-off. This implies that the power of market promotes healthy development of NPLs' management, disposal, and recovery.

3.5. Region

Some of the provinces are engaged into in the model, such as Tianjin, Fujian, and Hebei. Combined with the result of Stepwise discrimination, the factor of "province" has a relative small effect in classification. Further, when using the other significant variables except "province", the result of the classification has little change. The results of classification are well stated in the diagram below. In combination of Figure 6, we can do more in-depth analysis.

Ignoring the provinces with very few samples (Beijing, Hainan and Guangdong) in Figure 6 , the ratio of NPLs with zero recovery is relatively high, coming between 40% and 60%, provided that the obligors are located in Liaoning, Jiangxi, Shandong and Guangxi province. These provinces, other than Shandong province, are less developed regions, especially Jiangxi and Guangxi. The ratio of NPLs with zero recovery is relatively high in Shandong province because economic development is unbalanced there. Further analysis
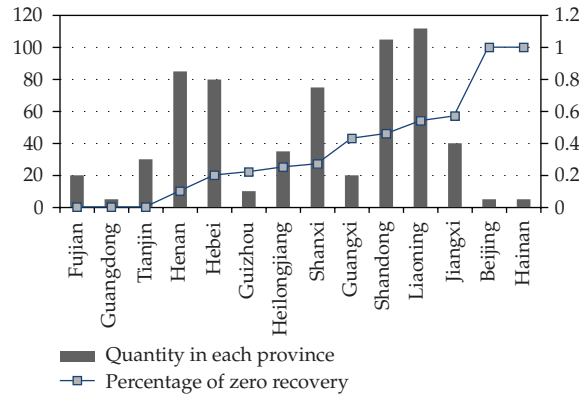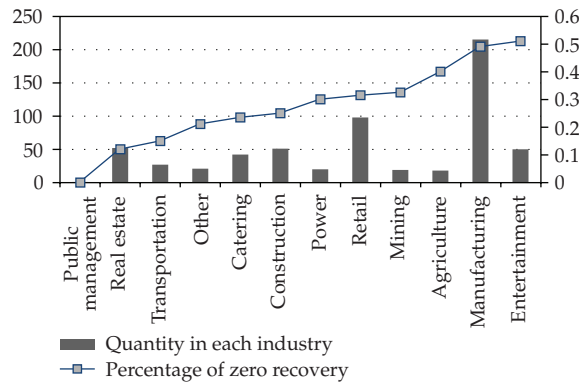
Figure 6

Figure 7

finds that most of the NPLs with zero recovery are from less developed cities. This may lead to the high ratio of NPLs with zero recovery.

The two highest ratios of NPLs with positive recovery are from Fujian and Tianjin. In these two provinces, NPLs commonly have positive recovery because both Fujian and Tianjin are developed regions in China. Less developed Henan province also has a relatively low ratio of NPLs with zero recovery, though it has not been referred in the model. A further analysis of the sample finds that most of the NPLs in Henan belong to subcategory and doubt-category loans and are bought from commercial banks. This is why the ratio of NPLs in Henan with positive recovery is high and variable, "Henan" is not significant in Stepwise discrimination model. Compared with region factors, the quality of loans plays a major role in determining whether they have recovery. To sum up, the economic conditions in the region surely influence the recovery of NPLs. Generally speaking, the more developed their economy is, the larger the probability of positive recovery is. Nevertheless, the factors about loan quality, such as the five-category classification of loans and the business situation of companies etc., play a more important role than region factor in determining whether loans are with positive recovery.

3.6. Industry

The results of Stepwise discrimination show that variables of industry have a small effect on the classification; only real estate industry and retail industry show a significant impact.

Table 3: The prediction results of whole samples.

			Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
Whole sample	Positive recovery	Bayesian Discrimination	406	19	95.5%
		Distance	395	30	92.9%
	Zero recovery	Bayesian	19	181	90.5%
		Distance	14	186	93.0%
	Total accuracy		Bayesian		93.9%
			Distance		93.0%

Figure 7 shows that the ratio of NPLs with zero recovery in real estate is approximate 13%, which implies that most of NPLs in the real estate industry have positive recovery. This is in line with the actual situation in real estate industry, which has been developing fast over the past 10 years in China. The other industries of which the ratios are also low are not significant. This is mainly because the effect of industry on classification has been categorized into the business situation and five-category classification, as concluded by analyzing corresponding samples in these industries.

4. Predicting Models and Accuracy of Prediction

Business situation, five-category classification, collateral state, province, and industry have been proved to be important in determining whether recovery rate of NPLs is zero or positive, according to discriminant function and data samples. In the sequel, we will use these variables to establish some discrimination models and verify the predicting competence of these models. Both Bayesian discrimination and Distance discrimination will be used to build predicting models with the variables selected by the Stepwise discrimination.

The most commonlyused method of modeling is to build models with the whole sample and then compare the difference between the predicted data and the actual amount; however, the adaptability of model established in this way needs to be considered carefully. The model based on the whole samples may be well adapted to the original samples but may also act badly for new samples. This is the so-called "over-fitting" problem. In order to solve this problem, the cross-validation method is employed. The whole sample is then divided into five equally subsamples. Each subsample has 125 observations which consist of 85 observations with positive recovery and 40 with zero recovery. Four of the five subsamples are adopted to fit a model each time, and the remained one would be used for prediction. The mean accuracy of the five submodels would be considered as the standard to select the best model. In this paper, the predicting efficiency is compared between Bayesian discrimination and Distance discrimination. The reason for the sample to be divided in this way lies in consideration of balance of sample size, computing complexity and the consistency of the statistical properties.

The results of the prediction out of whole samples by Bayesian discrimination and Distance discrimination are shown in Table 3 .

The prediction results using cross-validation in the two discriminating methods are shown in the following tables.

Table 4: The prediction results of the first model using cross-validation.

			Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
The first model	In-Sample	Positive recovery	Bayesian	323	17	95.0%
			Distance	319	21	93.8%
		Zero recovery	Bayesian	14	146	91.3%
			Distance	7	153	95.6%
		Total accuracy	Bayesian			93.8%
			Distance			94.4%
				Positive recovery (predicted)	Zero recovery (predicted)	
	Out-of-Sample	Positive recovery	Bayesian	77	8	90.6%
			Distance	72	13	84.7%
		Zero recovery	Bayesian	5	35	87.5%
			Distance	5	35	87.5%
		Total accuracy	Bayesian			89.6%
			Distance			85.6%

Table 5: The prediction results of the second model using cross-validation.

			Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
The second model	In Sample	Positive recovery	Bayesian	323	17	95.0%
			Distance	321	19	94.4%
		Zero recovery	Bayesian	9	151	94.4%
			Distance	8	152	95.0%
		Total accuracy	Bayesian			94.8%
			Distance			94.6%
				Positive recovery (predicted)	Zero recovery (predicted)	
	Out-of-Sample	Positive recovery	Bayesian	75	10	88.2%
			Distance	73	12	85.9%
		Zero recovery	Bayesian	1	39	97.5%
			Distance	1	39	97.5%
		Total accuracy	Bayesian			91.2%
			Distance			89.6%

Some further results can be calculated based on the results given in the above tables.

In Bayesian discrimination, the total mean in-sample accuracy of the five models is 94%, with 94.7% accuracy for positive recovery and 92.5% accuracy for zero recovery; the mean total out-of-sample accuracy of the five models is 92.64% with 93.6% accuracy for positive recovery and 90.5% accuracy for zero recovery. While in Distance discrimination,

Table 6: The prediction results of the third model using cross-validation.

			Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
TheThird model	In-Sample	Positive recovery	Bayesian	317	23	93.2%
			Distance	316	24	92.9%
		Zero recovery	Bayesian	10	150	93.8%
			Distance	9	151	94.4%
		Total Accuracy	Bayesian			93.4%
			Distance			93.4%
				Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
	Out of Sample	Positive recovery	Bayesian	83	2	97.6%
			Distance	81	4	95.3%
		Zero recovery	Bayesian	2	38	95.0%
			Distance	2	38	95.0%
		Total Accuracy	Bayesian			96.8%
			Distance			95.2%

Table 7: The prediction results of the fourth model using cross-validation.

			Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
The fourth Model	In-Sample	Positive recovery	Bayesian	324	16	95.3%
			Distance	314	26	92.4%
		Zero recovery	Bayesian	17	143	89.4%
			Distance	10	150	93.8%
		Total accuracy	Bayesian			93.4%
			Distance			92.8%
				Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
	Out-of-Sample	Positive recovery	Bayesian	82	3	96.5%
			Distance	78	7	91.8%
		Zero recovery	Bayesian	2	38	95.0%
			Distance	1	39	97.5%
		Total accuracy	Bayesian			96.0%
			Distance			93.6%

the mean total in-sample accuracy of the five models is 93.88% with 93.59% accuracy for positive recovery and 94.5% accuracy for zero recovery; the mean total out-of-sample accuracy of the five models is 90.24% with 89.6% accuracy for positive recovery and 91.5% accuracy for zero recovery.

Table 8: The prediction results of the fifth model using cross-validation.

			Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
The Fifth model	In-Sample	Positive recovery	Bayesian	323	17	95.0%
			Distance	321	19	94.4%
		Zero recovery	Bayesian	10	150	93.8%
			Distance	10	150	93.8%
		Total accuracy	Bayesian			94.6%
			Distance			94.2%
				Positive recovery (predicted)	Zero recovery (predicted)	Accuracy
	Out-of-Sample	Positive recovery	Bayesian	81	4	95.3%
			Distance	77	8	90.6%
		Zero recovery	Bayesian	9	31	77.5%
			Distance	8	32	80.0%
		Total accuracy	Bayesian			89.6%
			Distance			87.2%

Based on the prediction results, we can conclude the following.

(1) The in-sample accuracy and out-of-sample accuracy are both fairly high regardless of the discrimination method employed.

(2) Bayesian discrimination has a higher total accuracy than Distance discrimination. It is related to the sample structure and the characters of the two discriminations. The accuracy of Bayesian discrimination is more likely to be affected by the prior probability than the Distance discrimination. Provided we have chosen the right prior probability, Bayesian method would have better accuracy.

5. Predicting Zero Recovery for Obligors with Several Loans

As mentioned in Section 2, obligors are divided into two types: obligors having one loan or obligors with several loans. In this section, we extend the models in the above section to classify the zero recovery and positive recovery for the obligors with several loans. The main idea is described as follows.

First, assume that whether one loan is zero recovery or not is independent from other loans of the obligor. To be specific, assuming one obligor has n loans: $A_1, A_2, \ldots, A_n$, we can determine whether the obligor's recovery is zero or not by the information of this customer and all his loans and by the linear discriminating function f1 and f2 we derive from the model. Assuming that the recovery state of the ith loan is I_i, which is a discrete variable with value 1 if the loan is positive recovery and value 0 otherwise, then, the recovery state of the obligor's is $I = \sum_{i=1}^{n} I_i$. If $I > 0$, then this obligor has positive recovery and zero otherwise.

Since the in-sample and out-of-sample prediction accuracy of Bayesian discrimination is better, we select Bayesian method to predict the classification. The result is displayed in Table 9 .

Table 9: The prediction results of obligors having several loans.

	Be judged to be positive recovery	Be judged to be zero recovery	Accuracy
Positive recovery	533	59	90.1%
Zero recovery	18	211	92.0%
Total accuracy			90.6%

As is shown in Table 9 , the total accuracy is quite high. There is no big difference for the accuracy of single loan case, which proves the efficiency of the single loan model. Moreover, the result indicates that there exists significant difference in the obligor's economic condition and the quality of the loans between the obligors with positive recovery and those with zero recovery, no matter the obligors borrow one loan or several loans. So, it is essential for commercial banks and asset management companies to pay close attention to obligors instead of loans when defaults happen.

6. Conclusion and Future Work

This paper uses Stepwise discrimination to select important factors which determine whether an NPL has recovery or not. Combined with statistical analysis, we give an in-depth analysis of the significant factors on the basis of discriminant function and then we employ Distance discrimination and Bayesian discrimination to develop several models. We test the models' forecasting efficiency, and the results show that these models can reach a high accuracy in prediction.

We find that there are significant differences between the NPLs with recovery and the NPLs without recovery, which are shown in business situation, five-category classifications for loans, transfer mechanism of loan, collateral, industry, and region, respectively. The business situation of obligors and the quality of loans play a major role in classification.

Since it is difficult to access to financial data, some important information such as value of the collateral as well as company's financial statement can not be included in our models. This will no doubt affect the prediction accuracy of models. As the data in LossMetrics are collected from AMCs, the financial data were intentionally deleted by the commercial banks when they transferred NPLs to AMCs with abusing excuse of business secret. Hence it is believed that the results are possible to be improved provided that the commercial banks' data are used for modeling.

It should be pointed out that the structure and characteristics of NPLs will gradually change along with the economic development and time. For example, the policy-off NPLs occurred under a given history circumstance, might not take place again. Thus the changes make the studies on NPLs' recovery model an everlasting job.
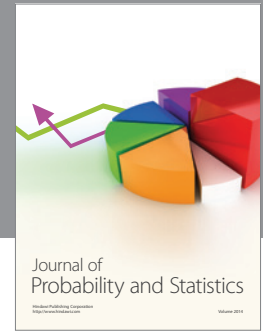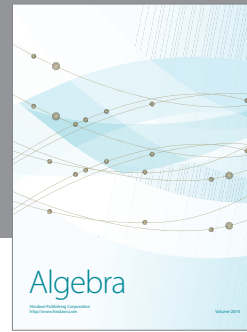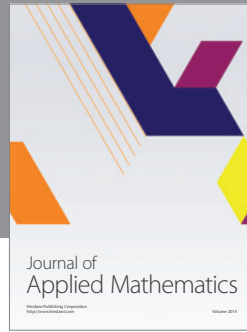
Acknowledgments

References

[1] T. Schuermann, "What do we know about Loss Given Default?" in *Credit Risk: Models and Management*, Working Paper, Risk, New York, NY, USA, 2nd edition, 2003.

[2] E. I. Altman and V. M. Kishore, "Almost everything you wanted to know about recoveries on defaulted bonds," *Financial Analysts Journal*, vol. 52, no. 6, pp. 57–64, 1996.

[3] V. V. Acharya, S. T. Bharath, and A. Srinivasan, "Understanding the recovery rates on defaulted securities," CEPR Discussion Papers 4098, Centre for Economic Policy Research, Washington, DC, USA, 2003.

[4] E. I. Altman, A. Resti, and A. Sironi, "Analyzing and explaining default recovery rates," ISDA Research Report, The International Swaps and Derivatives Association, London, UK, December 2001.

[5] L. V. Carty and D. Lieberman, "Defaulted bank loan recoveries," Moody's Special Report, Moody's Investors Service, New York, NY, USA, November 1996.

[6] G. M. Gupton, D. Gates, and L. V. Carty, "Bank-loan loss given default," Moody's Special Comment, Moody's Investors Service, New York, NY, USA, November 2000.

[7] L. Hurt and A. Felsovalyi, "Measuring loss on Latin American defaulted bank loans: a 27-year study of 27 countries," *The Journal of Lending & Credit Risk Management*, vol. 81, no. 2, pp. 41–46, 1998.

[8] G. M. Gupton, "Advancing loss given default prediction models: how the quiet have quickened," *Economic Notes*, vol. 34, no. 2, pp. 185–230, 2005.

[9] Z. Y. Chen, "Studies of loss given default," *Studies of International Finance*, no. 5, pp. 49–57, 2004 (Chinese).

[10] J. Wu, "Models of loss given default in IRB-research on core technology of New Capital Accord," *Studies of International Finance*, no. 2, pp. 15–22, 2005 (Chinese).

[11] H. F. Liu and X. G. Yang, "Estimation of loss given default: experience and inspiration from developed countries," *Journal of Financial Research*, no. 6, pp. 23–27, 2003 (Chinese).

[12] P. L. Shen, "The measurement of LCD in IRB approach," *Journal of Finance*, no. 12, pp. 86–95, 2005.

[13] X. K. Ye and H. L. Liu, "Research on structure of loss given default of NPLs in commercial bank," *Contemporary Finance*, no. 6, pp. 12–15, 2006 (Chinese).

[14] L. Bruch, *Discriminant Analysis*, Hafner Press, New York, NY, USA, 1975.

Advances in
Operations Research

Advances in
Decision Sciences

Journal of
Applied Mathematics

Algebra

Journal of
Probability and Statistics

The Scientific
World Journal

International Journal of
Differential Equations

International Journal of
Combinatorics

Advances in
Mathematical Physics

Submit your manuscripts at
http://www.hindawi.com

Hindawi

Journal of
Complex Analysis

Journal of
Mathematics

Mathematical Problems
in Engineering

Abstract and
Applied Analysis

Discrete Dynamics in
Nature and Society

International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces

International Journal of
Stochastic Analysis

Journal of
Optimization