Research Article

Data-Driven Robust Credit Portfolio Optimization for Investment Decisions in P2P Lending

Guotai Chi, Shijie Ding, and Xiankun Peng

Faculty of Management and Economics, Dalian University of Technology, Dalian 116024, China

Correspondence should be addressed to Shijie Ding; ding0601@126.com

Received 24 October 2018; Accepted 24 December 2018; Published 2 January 2019

Academic Editor: Emilio Gómez-Déniz

Copyright © 2019 Guotai Chi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-to-Peer (P2P) lending has attracted increasing attention recently. As an emerging micro-finance platform, P2P lending plays roles in removing intermediaries, reducing transaction costs, and increasing the benefits of both borrowers and lenders. However, for the P2P lending investment, there are two major challenges, the deficiency of loans' historical observations about the certain borrower and the ambiguity problem of estimated loans' distribution. In order to solve the difficulties, this paper proposes a data-driven robust model of portfolio optimization with relative entropy constraints based on an “instance-based” credit risk assessment framework. The model exploits a nonparametric kernel approach to estimate P2P loans' expected return and risk under the condition that the historical data of the same borrower is unavailable. Furthermore, we construct a robust mean–variance optimization problem based on relative entropy method for P2P loan investment decision. Using the real-world dataset from a notable P2P lending platform, Prosper, we validate the proposed model. Empirical results reveal that our model provides better investment performances than the existing model.

1. Introduction

Peer-to-peer lending, as an emerging online micro-finance, provides services that bring borrowers and lenders together virtually and help them to lend to and borrow from each other directly. P2P lending platforms play roles in removing traditional financial intermediaries, reducing transaction costs, and increasing the benefits of both borrowers and lenders; therefore, they improve the efficiency of financial market. However, due to the absence of traditional financial intermediaries which can use collateral, certified accounts, and other means to enhance the creditworthiness of borrowers, the information asymmetry between borrowers and lenders severely exist and the credit risk of P2P loan investment is very high.

Credit risk of P2P lending refers to the potential monetary loss arising from the default of a borrower to a loan. Efficient and reasonable investment in P2P loans needs to be based on the reliable credit risk distribution assessment. It is very challenging to estimate the credit risk distribution of P2P loans for the difficulty of obtaining the historical returns (or losses) data of the loan waiting for investment. In other words, the historical yield data about the same borrower is usually unavailable. Moreover, even the distribution of loans’ returns (or losses) is approximated from the limited available data or the expert knowledge, the approximation is usually not accurate, and it is also known as the distribution ambiguity (probability measure uncertainty) problem. In this paper, we formulate a data-driven robust portfolio optimization model based on an “instance-based” credit risk assessment method for investment decisions in P2P lending.

To help personal lenders mitigate the risk, the current online P2P lending platforms have taken some risk-reducing measures, such as filtering out the high-risk borrower whose FICO score is lower than a threshold, making a preliminary rating on each loan and providing investors with risk level of each loan. Thus, each loan is marked as a grade, like AA, A, B, C, D, E, or NR, and the loans with the same grade are considered to have the same risk level. These rating-based models are more suitable for traditional banks and lending institutions, since they have the capability to grant large amounts of loans to diversify their investments. However, the
individual investors just possess small amount of funds; they need more refined risk assessment methods and investment strategies.

Similar to bond investment, P2P investors can fund a portion, not the whole, of each loan. Therefore, investors can decide which loans to invest and, meanwhile, determine the amount of investment for each loan. This mechanism allows investors to construct a credit portfolio to mitigate risk.

Markowitz [1] proposes the famous mean-variance model, which is still widely used in portfolio selection and risk management. From then on, researchers propose a variety of mean-risk models, such as mean-downside risk model [2], mean-VaR model [3], mean-CVaR model [4], and so on. In practice, the distribution of the assets needs to be estimated firstly, and then the optimal portfolio can be identified by the optimization model.

For P2P lending investment, as mentioned above, such procedures face at least two major challenges, i.e., the deficiency of loans’ historical observations and the ambiguity problem of estimated loans’ distribution (probability measure uncertainty problem). Thus, this paper proposes a data-driven robust model of portfolio optimization based on relative entropy constraints combined with an instance-based credit risk assessment method.

Specifically, we use the “instance-based” credit risk assessment method proposed by Guo et al. [5] to evaluate the return and risk of each loan without sufficient historical data of loans for each individual borrower. In this instance-based framework, the expected return of each loan is predicted as a weighted average of historical loans of other similar borrowers, where the optimal weights are learnt based on kernel regression. Furthermore, using the moment information (mean and variance) of the new loans, we formulate the robust portfolio optimization model with relative entropy constraints, which could obtain an optimal portfolio under the worst scenario and has the ability of reducing the potential loss caused by the uncertainty of loans distribution.

Our work is somewhat related to the paper by Guo et al. [5] and the paper by Yam et al. [6]. Guo et al. [5] introduce the instance-based framework into credit risk assessment of P2P loan and use the classical mean-risk model to obtain the optimal allocation. Yam et al. [6] derive a robust mean-variance optimization model with relative entropy constrains on the uncertainty of the interaction between the returns of different assets and discuss its mathematical and financial properties in portfolio selection. Although some other scholars have contributed novel insights into credit risk assessment of P2P lending and robust optimization, to the best of our knowledge, few have taken both into consideration synthetically. The main contribution of this paper is that we propose a data-driven robust portfolio optimization model based on relative entropy constraints combined with instance-based risk assessment framework for P2P loan investment and obtain superior performance in numerical experiments.

The rest of this paper is organized as follows. Section 2 provides the literature review. Section 3 introduces the instance-based model for credit risk assessment, as well as the mathematical framework of kernel regression approach. In Section 4, we elaborate the robust optimization model based on relative entropy method and formulate a robust mean-variance optimization model for P2P lending investment. The empirical results on the effectiveness of our model is reported in Section 5. Finally, Section 6 concludes this work.

2. Literature Review

In order to assess risk and assist investment decisions making in P2P lending, researchers have done many studies: Emekter et al. [7] explore the dominated factors that explain the funding success and credit risk and, meanwhile, measure the performance of P2P loans. They find that credit grade, debt-to-income ratio, FICO score, and revolving line utilization play an important role in loan defaults; furthermore, loans with lower credit grade and longer duration may result in high mortality rate and higher interest rates charged on low credit grade borrowers are not sufficient to cover the potential loss for the higher likelihood of loan defaults. Thus, the authors suggest that investors should invest more to high grade loans. Similarly, Berkovich’s [8] study finds that high quality loans offer excess return.

The above researches investigate the factors determining the credit risk and analyze the performance of P2P loans; however, they do not propose a mechanism which assist individual investors in allocating loans effectively and making optimal investment decisions.

To help personal lenders mitigate the risk, the popular online P2P platforms, like Lending Club and Prosper, have developed credit scoring systems to assess the creditworthiness of each borrower based on data mining or machine learning techniques. There is a large body of existing literatures concerned with credit rating using data mining techniques, for example, linear discriminate analysis (LDA) [9], k-nearest neighbors [10], logistic regression [11], classification and regression trees (CART) [12], Markov chains [13], survival analysis [14], artificial neural network (ANN) [15], genetic methods [16], support vector machine (SVM) [17, 18], lasso-probit [19], and so on.

In the portfolio selection problem, full knowledge of the assets’ distribution is usually assumed to determine the optimal portfolio. In most real-life applications, we need to approximate the assets’ distribution. However, the approximations are not necessarily accurate, and it is known as the distribution ambiguity (probability measure uncertainty) problem.

The robust optimization algorithm is an attractive way to solve the portfolio selection problem under distribution ambiguity. As the exact parameters are unavailable, Natarajan et al. [20] use a set of parameters (which represent different distributions or scenarios) rather than a point estimation of the parameters to formulate the asset allocation problem. Following this idea, there are different ways to model ambiguity by using a set of parameters. Chen et al. [21] take the lower partial moments and CVaR as two risk measures and consider a tight bound which are likely to cover the possible parameters. Epstein [22] considered intervals that may include the actual parameters. Natarajan et al. [23] use a piecewise-linear concave utility function to derive
accurate and estimated optimal strategies for the expected utility model in the portfolio optimization issue under the worst-case scenarios. Paç and Pinar [24] use an ellipsoidal uncertainty set to represent the distribution ambiguity to identify the optimal portfolio.

Since relative entropy has the ability to measure the difference between two probability distributions (probability measures), it can be used to construct the uncertainty set for robust optimization. In the studies of Hansen and Sargent [25] and Calafora [26], relative entropy is used to model uncertainty and obtain the optimal investment decision. Yam et al. [6] derive a robust mean-variance optimization model with relative entropy constrains on the uncertainty of the interaction between the returns of different assets and discuss its mathematical and financial properties in portfolio selection.

In recent years, research on data-driven methods has been well studied. In this framework, it is assumed that investors only possess the information about history data of asset return. Bertsimas et al. [27] use KS test, χ² test, Anderson-Darling test, and some other testing tools to construct uncertainty sets and take the worst case of each set to formulate the robust optimization. They assume that the uncertainty sets are defined by certain structures and sizes based on the data points available. While the structure of uncertainty set in our study is not predefined, we consider the uncertainty of mean, covariance, and distribution synthetically. Kang et al. [28] propose a data-driven robust mean-CVaR portfolio selection model under the condition of distribution ambiguity and adopt a nonparametric bootstrap approach to calibrate the levels of ambiguity. Their work is based on the mean-CVaR framework with data of stock indices, while our work is based on the mean-variance framework with data of P2P loans.

3. Instance-Based Model for Credit Risk Assessment

Using historical data to evaluate future performance and potential loss is a convention. However, unlike bonds or stocks investment, the historical yield data about the same P2P borrower is usually unavailable. Thus, the risk assessment of new loan is very challenging. In this section, we briefly introduce the instance-based credit risk assessment model proposed by Guo et al. [5].

3.1. Instance-Based Assessment Framework. In this instance-based assessment framework, the expected return of each loan is estimated as a weighted average of historical observations of other borrowers’ closed loans. Specifically, for a new loan \(i\), using \(n\) past loans, each with an historical return \(R_j\) (\(j = 1, 2, \ldots, n\)), we can calculate the expected return of loan \(i\), \(\mu_i\), based on a weighted average of past loans’ actual returns:

\[
\mu_i = \sum_{j=1}^{n} w_{ij} R_j, \tag{1}
\]

where \(w_{ij}\) denotes the weight of loan \(j\) for predicting the expected return of loan \(i\). The weight depends on the similarity between loan \(i\) and loan \(j\). Intuitively, the more the similarity, the greater the weight. The calculation of the weight will be introduced in Section 3.2.

The weighted returns of the past loans are assumed as historical observations of a new loan. According to this line of thought, taking variance as the risk measure, weighted variance of past loans are used to assess the new loan’s risk, that is,

\[
\sigma_i^2 = \sum_{j=1}^{n} w_{ij} (R_j - \mu_i)^2, \tag{2}
\]

where \(w_{ij}\), \(R_j\), and \(\mu_i\) have the same meanings as (1).

The absolute deviation between two loans’ default probabilities is used to measure the similarity; the smaller the absolute deviation, the more the similarity, and, therefore, the larger the weight. In particular, absolute deviation of default probabilities between loans \(i\) and \(j\) is defined as follows: \(d_{ij} = |p_i - p_j|\), where \(p_i\) and \(p_j\) are the default probabilities of loans \(i\) and \(j\), respectively. Kernel regression is exploited to investigate the nonlinear relationship between the absolute deviation and the weight. This process will be introduced in the next subsection.

3.2. Kernel Regression of Return and Risk. Kernel regression is a nonparameter statistical method to investigate the nonlinear relation between random variables, which is based on the kernel density estimation. First of all, the preliminaries of kernel estimation are introduced.

Given \(n\) realizations \(z_j, j = 1, \ldots, n\), of random variable \(z\), the kernel estimation \(\hat{p}(z)\) of the probability density function \(p(z)\) is defined by

\[
\hat{p}(z) = \frac{1}{nh} \sum_{j=1}^{n} K \left( \frac{z - z_j}{h} \right), \tag{3}
\]

where \(K(\cdot)\) is a kernel function and \(h\) is a smoothing parameter.

**Kernel function** \(K(\cdot)\) is nonnegative and bounded and, meanwhile, satisfies the following properties:

(a) \(\int_{-\infty}^{\infty} K(z)dz = 1\); (b) \(\int_{-\infty}^{\infty} zK(z)dz = 0\); (c) \(\int_{-\infty}^{\infty} z^2 K(z)dz < \infty\).

There are a range of commonly used kernel functions, such as uniform, triangular, biweight, triweight, and Gaussian [29]. Because the kernel estimation is insensitive to the choice of kernel function, we use the Gaussian kernel function due to its convenient mathematical properties, which is written as

\[K(z) = \left( \frac{1}{\sqrt{2\pi}} \right) e^{-z^2/2}.\]

The smoothing parameter \(h = h(n)\) is also called the bandwidth that depends on the sample size \(n\). Specifically, \(h(n)\) and \(n \cdot h(n)\) decrease to 0 as \(n\) tend to \(\infty\).

Many literatures reveal that the choice of kernel function does not affect the estimation significantly; however, the choice of the bandwidth is a vital issue [30, 31]. The determination of the bandwidth will be shown in detail in Section 5.3.

In the following, we introduce the kernel regression model proposed by Nadaraya [32]. Theoretically, we assume that each observation is denoted as \((X, Y)\) which is a random
vector $\mathbb{R}^2$-valued. With the sample set, $\{(x_j, y_j)\}_{j=1}^n$, the kernel estimator $\hat{y}$ of the target $y$ given its predictive observation $x$ is defined as

$$
\hat{y} = \sum_{j=1}^n \left[ \frac{K\left(\frac{x - x_j}{h}\right)}{\sum_{j=1}^n K\left(\frac{x - x_j}{h}\right)} \cdot y_j \right],
$$

where $K(\cdot)$ is a kernel function and $h$ is the bandwidth.

For the instance-based credit risk modeling, the set of historical observations is represented as $\{(p_i, R_i)\}_{i=1}^n$, where $p_i$ and $R_i$ are the default probability and return rate of the $i$th loan, respectively. Thereby, the estimation of the $i$th loan’s return could be written as

$$
\hat{\mu}_i = \sum_{j=1}^n \left[ \frac{K\left(\frac{p_i - p_j}{h}\right)}{\sum_{j=1}^n K\left(\frac{p_i - p_j}{h}\right)} \cdot R_j \right],
$$

Note that the determination of loans’ default probability will be introduced in Section 5.1.

Comparing (1) to (5), we can represent the optimal weight $w_{ij}$ as

$$
w_{ij} = \frac{K\left(\frac{p_i - p_j}{h}\right)}{\sum_{j=1}^n K\left(\frac{p_i - p_j}{h}\right)}. 
$$

Using the optimal weight $w_{ij}$ and the expected return $\hat{\mu}_i$ derived from (5), (2) can be rewritten as

$$
\hat{\sigma}_i^2 = \sum_{j=1}^n \left[ \frac{K\left(\frac{p_i - p_j}{h}\right)}{\sum_{j=1}^n K\left(\frac{p_i - p_j}{h}\right)} \cdot (R_j - \hat{\mu}_i)^2 \right].
$$

### 4. Robust Investment Decision Model

Similar to bond investment, P2P lenders can invest a portion of each loan. Thus, P2P loan investment decisions can be transformed into a credit portfolio optimization problem. This section introduces the portfolio optimization model for investment decisions in P2P lending, which accounts for the uncertainty of the distribution of the loans. We start from the classical mean-variance optimization model proposed by Markowitz [1] to its tractable robust counterpart.

#### 4.1. Robust Optimization Model Based on Relative Entropy Constraints

In the classical mean-variance optimization model, the optimal asset allocation strategy is identified by solving the tradeoff between risk and return according to investors’ risk preference. A portfolio that invests in $n$ assets is represented as a vector of weights, $\lambda \in \mathbb{R}^n$, where each weight denotes the proportion of wealth allocated to an asset. Then the return and risk of the portfolio become $\lambda^T \mu$ and $\lambda^T \Sigma \lambda$, respectively, where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ are the expected return and the covariance matrix of the assets’ returns under the probability measure (or probability distribution) $P$, respectively. Here, $P$ represents the ideal estimated market condition where $\mu$ and $\Sigma$ estimated by using all available information, including historical observations, news, expert knowledge, and so on, are assumed as the actual expected return and covariance matrix. Thus, the classical mean-variance portfolio selection problem (MV) can be formulated as

$$
\text{(MV)}: \min_{\lambda} \lambda^T \Sigma \lambda \\
\text{s.t. } \lambda^T \mu \geq R^*, \\
\lambda \in \Omega,
$$

where $\Omega \subseteq \mathbb{R}^n$ denotes the set of feasible portfolios and $R^*$ is the required return rate specified by the investor.

In reality, the assumption that the expected return $\mu$ and covariance matrix $\Sigma$ are known with certainty is less reasonable. It is quite possible that the estimated parameters are different from the actual ones. Thus, the optimal portfolio identified by using the estimated inputs parameters $\mu$ and $\Sigma$ directly may be inappropriate. Robust optimization seeks for portfolios that are insensitive to the uncertain in the parameters and the solutions that must be feasible no matter what the actual value of the parameters is.

The investors might consider a set of probability measures, i.e., an uncertainty set, to cover a range of scenarios based on their assessments, and then use robust optimization to obtain approximate optimal strategies for the worst scenarios within the uncertainty set. In this paper, we define $\mathcal{Q}$ as the set of probability measures representing the possible scenarios, $\mu_Q$ and $\Sigma_Q$ as the expected return and covariance matrix estimated under the probability measure $Q \in \mathcal{Q}$. Mathematically, the robust counterpart of the classical mean-variance optimization problem (RMV) can be written as

$$
\text{(RMV)}: \min_{\lambda} \sup_{Q \in \mathcal{Q}} \lambda^T \Sigma_Q \lambda \\
\text{s.t. } \inf_{Q \in \mathcal{Q}} \lambda^T \mu_Q \geq R^*, \\
\lambda \in \Omega,
$$

It is rational to assume that the actual value of the parameters is in the neighborhood of the estimator. Thus, we can generate the uncertainty set $\mathcal{Q}$ based on the assumption that the measures in the set should be not far from the ideal measure $P$. Relative entropy, also known as the Kullback–Leibler divergence, can be used to measure the difference between probability measures. The relative entropy of the measure $Q$ in $\mathcal{Q}$ with respect to the measure $P$ is

$$
D_{KL}(Q \parallel P) := \int q(x) \ln \frac{q(x)}{p(x)} dx,
$$

where $p(x)$ and $q(x)$ are the probability density functions (pdf) of the loans’ returns under probability measures $P$ and $Q$, respectively. In the context of mean-variance analysis, relative entropy $D_{KL}(Q \parallel P)$ can be rewritten as

$$
D_{KL}(Q \parallel P) = \frac{1}{2} \left[ \ln |V| - \ln |V_Q| + \text{tr} \left( V^{-1} V_Q \right) - n \\
+ (\mu - \mu_Q)^T V^{-1} (\mu - \mu_Q) \right],
$$

where $|\cdot|$ denotes the determinant of a matrix, $\text{tr}$ denotes the trace of a matrix, $V$ and $V_Q$ are the variance and covariance matrices of the loans’ returns under the probability measure $P$ and $Q$, respectively, and $n$ is the number of assets.
where $\mu$, $V$, $\mu_Q$, and $V_Q$ carry the same meaning as in (8) and (9); \(\text{tr}(V)\), \(|V|\), and $V$ be the trace, the determinant, and the transpose of $V$, respectively; $n$ is the amount of assets in the portfolio.

Let $\mathcal{U}$ denote the set of parameters $(\mu_Q, V_Q)$ under the measure $Q$ in $\mathcal{F}$. Using the constraint of relative entropy, we can rewrite the robust optimization model, (9), as

\[
(RMV-RE): \min_{\lambda} \max_{(\mu_Q, V_Q) \in \mathcal{U}} \lambda^T V_Q \lambda \\
\text{s.t.} \quad \min_{(\mu_Q, V_Q) \in \mathcal{U}} \lambda^T \mu_Q \geq R^*, \\
D_{KL} (Q \parallel P) \leq K, \\
\lambda \in \Omega.
\]

where $K$ is a positive constant and determines the size of uncertainty set. Parameter $K$ measures the level of uncertainty and reflects the investors' confidence in $\mu$ and $V$ estimated under probability measure $P$, i.e., the greater $K$'s value, the less confidence.

Yam et al. [6] prove that the robust mean-variance portfolio selection model based on relative entropy method (RMV-RE) can be formulated as quadratic optimization problem, which is a tractable formulation and can be efficiently solved. That is,

\[
\min_{\lambda \in \mathbb{R}^n} \lambda^T V_* \lambda \\
\text{s.t.} \quad \lambda^T \mu_* \geq R^*, \\
\lambda \in \Omega.
\]

Herein, $\mu_* = \zeta \mu$, $V_* = V + \zeta (1-\zeta) \mu \mu^T$ and $\zeta \in (0, 1]$ is related to $K$ in (12) closely, which reflects the level of confidence in $\mu$ and $V$ estimated under measure $P$. For example, $\zeta=1$ means that investors believe the estimated $\mu$ and $V$ are the true parameters. And as $\zeta$ decreases, the investor’s confidence is weaker. The details of the proof are referred to by Yam et al. [6].

4.2. Robust Mean-Variancne Portfolio Optimization Model in P2P Lending. In the Section 3.2, we estimated each loan’s expected return and variance of return, i.e., $\bar{\mu}$ and $\bar{\sigma}$, using the instance-based credit risk assessment model. Let $\bar{\mu} = (\bar{\mu}_1, \bar{\mu}_2, ..., \bar{\mu}_n)^T$ and

\[
\bar{\sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \sigma_n \end{bmatrix},
\]

denote the expected return vector and the covariance matrix of the loans’ returns under the probability measure $P$. Here, we assume that the correlation between P2P loans is negligible. Now we can rewrite (13) as

\[
\min_{\lambda \in \mathbb{R}^n} \lambda^T (\bar{\mu} + \zeta (1-\zeta) \bar{\mu} \bar{\mu}^T) \lambda \\
\text{s.t.} \quad \lambda^T (\zeta \bar{\mu}) \geq R^*, \\
\lambda \in \Omega.
\]

The feasible region $\Omega$ of our problem is defined by the following constraints:

(1) The value of the portfolio remains at its initial value, i.e., $\sum_i \lambda_i = 1$.
(2) Short-selling is forbidden; thus $\lambda_i \geq 0$.
(3) For each loan, the amount that lender can invest is no more than the borrower request, $m_i$; thereby, $\lambda_i M \leq m_i$, where $M$ is the total investment amount and investor has available.

5. Empirical Analysis

In this section, we investigate the validity of the robust mean-variance portfolio optimization model in P2P lending using the real-world dataset from a notable P2P lending platform, Prosper. All numerical experiments are performed by using MATLAB on PC.

5.1. Data Description and Preprocess. The dataset for empirical study is from a notable P2P lending platform in the United States, Prosper. It consists of 17,001 loans including 3039 default loans and 13908 completed loans, whose issue dates within the period from November 2005 to March 2014.

Using the data, a credit scoring model is learnt to transform the loan attributes into the default probability. The loan attributes are as follows: the borrower’s FICO score which reflects borrower’s creditworthiness, the borrower’s number of inquiries in the past six months, the monetary amount of the loan, the homeownership status of the borrower, the debt-to-income ratio of the borrower, the borrower’s current delinquencies representing the number of accounts delinquent, and the borrower’s number of public records in the past 10 years (Row 1-7 in Table 1). The target variable is a binary variable (0 represents completed and 1 represents default), as described in Row 8 of Table 1.
There exist many credit scoring models to predict the default probability of a loan, such as: Xgboost model [33–35], hybrid KMV model [36], credit scoring based on genetic algorithms [37, 38], and so on. However, discussing how to choose and construct the optimal credit scoring model is beyond the scope of this study, and we use the most popular model, logistic regression, to make the prediction in this preprocessing step.

We randomly divide the dataset into two parts, one containing 40% of all loans for determining the optimal bandwidth \( h \) in (5), which will be described in detail in Section 5.3 and the second part containing 60% of the loans. Moreover, using k-fold cross-validation, we randomly divide the second part into 20 subsets, each of which contains approximately 510 loans. In each round, one of the subsets is used as the testing set which consists of loans waiting to be invested, thus their pay-back statuses are unknown, and all other subsets are taken as a training set which consists of historical loans with known yield.

### 5.2. Model Description

In this paper, we propose a robust credit portfolio optimization model for investment decisions in P2P lending. In order to show its effectiveness, we compare it with a benchmark model proposed by Guo et al. [5]. In the following, we describe models in detail.

IOM is the instance-based model proposed by Guo et al. [5]. Each loan is assessed using kernel weights and the historical performance of similar loans. Then use the classical mean-variance model (8), to identify the optimal allocation strategy. The performance of this model outperforms some rating-based models, as the results of Guo et al. [5] show.

RIOM is the robust instance-based model in this study. Expected return and risk of each loan are also assessed based on the “instance-based” assessment framework. However, we use the robust model of credit portfolio optimization based on relative entropy method, Equation (15), to obtain the optimal investment decision.

We compare the two models by the following procedure:

1. Train the credit risk assessment model with the training set, and use the trained model to predict the expected return \( \mu_i \) and variance \( \sigma_i \) of each loan in the testing set. Thus, the expected return vector and the covariance matrix, \( \mu \) and \( V \), can be obtained.

2. For each model, feed the predicted expected return vector \( \mu \) and the covariance matrix \( V \) of the testing loans into the portfolio optimization algorithm, and compute the performance of investment on the optimal portfolio.

3. Compare the return rate of the two models.

### 5.3. Analysis of Results

As mentioned before, we select the Gaussian kernel, \( K(\zeta) = 1/\sqrt{2\pi} e^{-\zeta^2/2} \), as the kernel function. And the important parameter in the kernel regression model, bandwidth \( h \), is optimized by the following leave-one-out cross validation:

\[
h_{optimal} = \arg \min_h CV(h)
\]

\[
= \arg \min_h \sum_{i=1}^{n} (\mu_h(p_{-i}) - \mu_i)^2,
\]

where \( \mu_h(p_{-i}) \) is the leave-one-out estimation of expected return rate \( \mu_i \), specifically,

\[
\mu_h(p_{-i}) = \frac{\sum_{j=1,j\neq i}^{n} K\left( \frac{p_i - p_j}{h} \right) \cdot R_j}{\sum_{j=1,j\neq i}^{n} K\left( \frac{p_i - p_j}{h} \right)}.
\]

The curve of \( CV(h) \) is exhibited in Figure 1. The shape of the curve clearly shows a minimal point and \( h \) corresponding to the minimal point is the optimal bandwidth for the model.

To apply the robust credit portfolio optimization method to obtain the optimal investment strategy in problems (13), we select the parameter \( \zeta = 0.75 \), the investment amount \( M = 15 \) thousand dollars, and the required rate of return \( R^* = 0.05 \). We also set the risk-free return rate as 0.025, which is about equivalent to the average yield of T-Bills over the same period. And we use the MATLAB built-in solver “quaprog” to solve the two portfolio optimization problems.

Table 2 summarizes investment return rate of each test subset and the average performance of the Prosper dataset. It shows that the two portfolios are almost always efficient and feasible, except subset 16. The results also show that the actual performances of the optimal portfolio derived from RIOM always outperform the optimal portfolio from IOM. And the Sharpe ratio shows that median-based optimal portfolio performs better as well.
The number of parameters set

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Subset & IOM & RIOM \\
\hline
1 & 0.0501 & 0.0566 \\
2 & 0.0550 & 0.0633 \\
3 & 0.0540 & 0.0618 \\
4 & 0.0564 & 0.0696 \\
5 & 0.0627 & 0.0714 \\
6 & 0.0543 & 0.0629 \\
7 & 0.0532 & 0.0688 \\
8 & 0.0605 & 0.0711 \\
9 & 0.0593 & 0.0706 \\
10 & 0.0546 & 0.0664 \\
11 & 0.0637 & 0.0701 \\
12 & 0.0567 & 0.0640 \\
13 & 0.0468 & 0.0569 \\
14 & 0.0519 & 0.0663 \\
15 & 0.0544 & 0.0620 \\
16 & 0.0357 & 0.0472 \\
17 & 0.0588 & 0.0710 \\
18 & 0.0607 & 0.0774 \\
19 & 0.0544 & 0.0655 \\
20 & 0.0625 & 0.0808 \\
\hline
Average & 0.0553 & 0.0662 \\
\hline
\end{tabular}
\caption{Rate of return from the optimal portfolio on the Prosper dataset.}
\end{table}

In order to test and verify that the conclusions obtained from the above experiments are stable, we consider different investment amounts and required returns as input parameters for portfolio selection and keep other conditions unchanged. As summarized in Table 3, we consider nine parameters pairs about required return rate \( R^* \) and investment amount \( M \).

The computational results for each parameters pair are summarized in Table 4. Table 4 shows performance comparison of the two optimal portfolios from the perspectives of actual return rate of investment. The more intuitive results are shown in Figure 2, which shows the actual return rate comparison of the two models. The first 9 numbers of the horizontal axis in Figure 2 represent the corresponding parameters combinations (sets 1 through 9 from Table 3), and the number 10 shows the average. We can find that the RIOM model outperforms the IOM model comprehensively.

In conclusion, the optimal portfolio identified from the robust optimization model in this study is more efficient than the existing model. And the performance of our model is more robust and stable.

### 6. Conclusions

In this paper, we formulate a data-driven robust model of portfolio optimization with relative entropy constraints based on an instance-based credit risk assessment framework for investment decisions in P2P lending. This P2P lending investment decision model has at least three advantages. Firstly, it provides a more refined measure of P2P loans’ risk and reveals a more intuitive and quantized risk estimate to investors, instead of just labelling each loan with a credit grade. Secondly, this model can estimate each loan’s expected return and risk when the historical observation of the same borrower is unavailable. Finally, this model considers the loans’ distribution ambiguity (probability measure uncertainty) problem and uses relative entropy to model parameter uncertainty to ensure the optimal allocation strategy efficient and feasible under various actual scenarios. Numerical experiments imply that the P2P lending investment decision model using the robust optimization with relative entropy constraints provides better performance than existing model.
Table 4: Investment performances of input parameters for portfolio selection.

<table>
<thead>
<tr>
<th>Subset</th>
<th>$R^2 = 5%$</th>
<th>$R^2 = 5.5%$</th>
<th>$R^2 = 6%$</th>
<th>$R^2 = 5%$</th>
<th>$R^2 = 5.5%$</th>
<th>$R^2 = 6%$</th>
<th>$R^2 = 5%$</th>
<th>$R^2 = 5.5%$</th>
<th>$R^2 = 6%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IOM</td>
<td>RIOM</td>
<td>IOM</td>
<td>RIOM</td>
<td>IOM</td>
<td>RIOM</td>
<td>IOM</td>
<td>RIOM</td>
<td>IOM</td>
</tr>
<tr>
<td>1</td>
<td>0.0598</td>
<td>0.0727</td>
<td>0.0601</td>
<td>0.0762</td>
<td>0.0502</td>
<td>0.0722</td>
<td>0.0501</td>
<td>0.0566</td>
<td>0.0589</td>
</tr>
<tr>
<td>2</td>
<td>0.0500</td>
<td>0.0592</td>
<td>0.0601</td>
<td>0.0779</td>
<td>0.0675</td>
<td>0.0851</td>
<td>0.0550</td>
<td>0.0633</td>
<td>0.0517</td>
</tr>
<tr>
<td>3</td>
<td>0.0441</td>
<td>0.0571</td>
<td>0.0491</td>
<td>0.0698</td>
<td>0.0735</td>
<td>0.0934</td>
<td>0.0540</td>
<td>0.0618</td>
<td>0.0598</td>
</tr>
<tr>
<td>4</td>
<td>0.0525</td>
<td>0.0602</td>
<td>0.0658</td>
<td>0.0903</td>
<td>0.0636</td>
<td>0.0754</td>
<td>0.0564</td>
<td>0.0696</td>
<td>0.0512</td>
</tr>
<tr>
<td>5</td>
<td>0.0532</td>
<td>0.0620</td>
<td>0.0631</td>
<td>0.0829</td>
<td>0.0513</td>
<td>0.0731</td>
<td>0.0627</td>
<td>0.0714</td>
<td>0.0566</td>
</tr>
<tr>
<td>6</td>
<td>0.0634</td>
<td>0.0747</td>
<td>0.0564</td>
<td>0.0762</td>
<td>0.0717</td>
<td>0.1105</td>
<td>0.0543</td>
<td>0.0629</td>
<td>0.0570</td>
</tr>
<tr>
<td>7</td>
<td>0.0613</td>
<td>0.0736</td>
<td>0.0547</td>
<td>0.0754</td>
<td>0.0551</td>
<td>0.0884</td>
<td>0.0532</td>
<td>0.0688</td>
<td>0.0528</td>
</tr>
<tr>
<td>8</td>
<td>0.0529</td>
<td>0.0594</td>
<td>0.0505</td>
<td>0.0616</td>
<td>0.0685</td>
<td>0.0858</td>
<td>0.0605</td>
<td>0.0711</td>
<td>0.0545</td>
</tr>
<tr>
<td>9</td>
<td>0.0548</td>
<td>0.0647</td>
<td>0.0550</td>
<td>0.0736</td>
<td>0.0559</td>
<td>0.0493</td>
<td>0.0593</td>
<td>0.0706</td>
<td>0.0507</td>
</tr>
<tr>
<td>10</td>
<td>0.0474</td>
<td>0.0574</td>
<td>0.0472</td>
<td>0.0634</td>
<td>0.0499</td>
<td>0.0794</td>
<td>0.0546</td>
<td>0.0664</td>
<td>0.0528</td>
</tr>
<tr>
<td>11</td>
<td>0.0597</td>
<td>0.0730</td>
<td>0.0602</td>
<td>0.0795</td>
<td>0.0661</td>
<td>0.1090</td>
<td>0.0637</td>
<td>0.0701</td>
<td>0.0562</td>
</tr>
<tr>
<td>12</td>
<td>0.0644</td>
<td>0.0768</td>
<td>0.0541</td>
<td>0.0673</td>
<td>0.0624</td>
<td>0.1042</td>
<td>0.0567</td>
<td>0.0640</td>
<td>0.0529</td>
</tr>
<tr>
<td>13</td>
<td>0.0635</td>
<td>0.0785</td>
<td>0.0709</td>
<td>0.0895</td>
<td>0.0532</td>
<td>0.0662</td>
<td>0.0468</td>
<td>0.0569</td>
<td>0.0637</td>
</tr>
<tr>
<td>14</td>
<td>0.0593</td>
<td>0.0744</td>
<td>0.0626</td>
<td>0.0751</td>
<td>0.0634</td>
<td>0.1204</td>
<td>0.0519</td>
<td>0.0663</td>
<td>0.0568</td>
</tr>
<tr>
<td>15</td>
<td>0.0523</td>
<td>0.0636</td>
<td>0.0485</td>
<td>0.0609</td>
<td>0.0571</td>
<td>0.0987</td>
<td>0.0544</td>
<td>0.0620</td>
<td>0.0577</td>
</tr>
<tr>
<td>16</td>
<td>0.0549</td>
<td>0.0705</td>
<td>0.0684</td>
<td>0.0893</td>
<td>0.0508</td>
<td>0.1264</td>
<td>0.0357</td>
<td>0.0472</td>
<td>0.0642</td>
</tr>
<tr>
<td>17</td>
<td>0.0549</td>
<td>0.0666</td>
<td>0.0549</td>
<td>0.0757</td>
<td>0.0538</td>
<td>0.0677</td>
<td>0.0588</td>
<td>0.0710</td>
<td>0.0674</td>
</tr>
<tr>
<td>18</td>
<td>0.0546</td>
<td>0.0629</td>
<td>0.0512</td>
<td>0.0615</td>
<td>0.0560</td>
<td>0.0610</td>
<td>0.0607</td>
<td>0.0774</td>
<td>0.0585</td>
</tr>
<tr>
<td>19</td>
<td>0.0492</td>
<td>0.0555</td>
<td>0.0572</td>
<td>0.0685</td>
<td>0.0657</td>
<td>0.0436</td>
<td>0.0544</td>
<td>0.0655</td>
<td>0.0434</td>
</tr>
<tr>
<td>20</td>
<td>0.0554</td>
<td>0.0645</td>
<td>0.0413</td>
<td>0.0504</td>
<td>0.0596</td>
<td>0.0366</td>
<td>0.0625</td>
<td>0.0808</td>
<td>0.0562</td>
</tr>
</tbody>
</table>

Average: 0.0554 0.0664 0.0732 0.0598 0.0823 0.0553 0.0662 0.0560 0.0731 0.0595 0.0807 0.0552 0.0663 0.0564 0.0729 0.0597 0.0819
Data Availability

The data this paper used is downloaded from the website of Prosper: https://www.prosper.com/invest/download.aspx.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research is supported by the National Natural Science Foundation of China (Grants nos. 71471027, 71731003, and 71873103), the National Social Science Foundation of China (Grant no. 16BVT017), National Natural Science Foundation of China Youth Project (Grant no. 16BJY016), Liaoning Economic and Social Development Key Issues (Grant no. 2015LSKS00ZD01-05), and Liaoning Provincial Social Science Planning Fund Project (Grant no. L16BJY016). The authors acknowledge the organizations mentioned above.

References


