A RISK-ORIENTED CLUSTERING APPROACH FOR ASSET CATEGORIZATION AND RISK MEASUREMENT

Lu Liu

A thesis submitted in partial fulfillment of the requirements for the
Master's degree in Master of Science in Management

Telfer School of Management
Master of Science in Management
University of Ottawa

A Risk-oriented Clustering Approach for Asset Categorization and Risk Measurement

Lu Liu

Master's Degree

Telfer School of Management
University of Ottawa

2019

Abstract

When faced with market risk for investments and portfolios, people often calculate the risk measure, which is a real number mapping to each random payoff. There are many ways to quantify the potential risk, among which the most important input is the features from future performance. Future distributions are unknown and thus always estimated from historical Profit and Loss (P&L) distributions. However, past data may not be appropriate for estimating the future; risk measures generated from single historical distributions can be subject to error. To overcome these shortcomings, one natural way implemented is to identify and categorize similar assets whose Profit and Loss distributions can be used as alternative scenarios.

In practice, one of the most common and intuitive categorizations is sector, based on industry. It is widely agreed that companies in the same sector share the same, or related, business types and operating characteristics. But in the field of risk management, sector-based categorization does not necessarily mean assets are grouped in terms of their risk profiles, and we show that risk measures in the same sector tend to have large variation. Although improved risk measures related to the distribution ambiguity has been discussed at length, we seek to develop a more risk-oriented categorization by providing a new clustering approach. Furthermore, our method can better inform us of the potential risk and the extreme worst-case scenario within the same category.

Contents

1. Introduction *9*

1.1 Background **9**

1.2 Literature Review **13**

1.3 Research Question **21**

1.4 Contribution and Outline **23**

2. Risk Measures *25*

2.1 The Theory of Convex Risk Measures **25**

2.2 Value at Risk **28**

2.3 Conditional Value at Risk **31**

2.4 Spectral Risk Measures **36**

3. Clustering *43*

3.1 Overview **43**

3.2 Distance Measures **45**

3.3 Clustering Algorithms **48**

3.3.1 Hierarchical Clustering 50

3.3.2 K-means Clustering 54

4. Methodology and Validation Approach *58*

4.1 Framework **58**

4.1.1 Wasserstein Metric and Its Connection to Risk Measures 58

4.1.2 Uncertainty Set Based on Wasserstein Distance and the Size 65

4.1.3 Wasserstein-based Clustering 66

Wasserstein-based Hierarchical Clustering 69

4.1.4 Method Summary 70

4.2 Validation Approach: Backtesting **70**

4.2.1 Value at Risk Backtesting 71

4.2.2 Conditional Value at Risk Backtesting 72

4.2.3 Spectral Risk Measures Backtesting 74

5. Results *76*

5.1 Data and Overview **76**

5.2 Numerical Results 78

5.2.1 Wasserstein-based Clustering Categorization 78

5.2.2 Spectral Risk Measure Comparisons among Three Categorizations 80

5.2.3 Verification of the Relationship between Wasserstein Distance and Spectral Risk Measures 86

5.2.4 Backtesting of Value at Risk, Conditional Value at Risk and Spectral Risk Measures 88

6. Conclusion and Direction of Future Research 95

Nomenclature

CVaR Conditional Value at Risk

P&L Profit and Loss

SRMs Spectral Risk Measures

VaR Value at Risk

List of Tables

1. Distance function in clustering

2. Wasserstein-based clustering result

3. VaR Backtesting Comparison

4. CVaR Backtesting Comparison

5. Spectral Risk Measures Backtesting Comparison

List of Figures

1. 90% VaR and CVaR of a hypothetical Profit & Loss distribution
2. R&L distributions with the same VaR but different CVaR
3. Exponential weighting functions $\phi(\gamma)$ when absolute risk aversion $k = 5$ and $k = 25$
4. Dendrogram for clustering solution
5. Wasserstein distance between two box masses
6. Wasserstein distance between two univariate distributions μ_1 and μ_2
7. Wasserstein distance ($r = 2$): example
8. Wasserstein distance ($r = 2$) and transportation cost: example
9. P&L distributions among sectors
10. SRMs comparison among 6 sectors
11. SRMs comparison among 6 clusters based on Euclidean distance clustering
12. SRMs comparison among 6 clusters based on Wasserstein distance clustering
13. 95% CVaR difference and Wasserstein distance
14. 95% CVaR difference and Euclidean distance

1. Introduction

1.1 Background

With the globalization of finance and the volatility of financial markets increasing, revolutionary changes have taken place in the theory and practice of global financial institution management. Financial risk management has then become the foundation and core of the management of modern financial institutions. Financial risk refers to the potential loss of funds due to uncertainty in economic activities and the possibility of loss. The main risks faced by financial institutions involve market risk, credit risk, liquidity risk, operational risk and political risk, etc.

Among all those financial risks, market risk and credit risk are the two most important candidates. In the past, in the context of relatively stable financial market prices, more attention is paid to the credit risk, and market risk is ignored to some extents. For example, financial risk management in the 1970s is almost entirely related to the management of credit risks.

Since the collapse of the Bretton Woods system in the early 1970s, changes in the prices of financial products such as exchange rates and interest rates have become increasingly frequent and disorderly under the floating exchange rate regime. The rapid development of financial innovation and information technology since the 1980s, as well as the trend of financial liberalization in the world, has led to a more volatile financial market. In addition, due to the need to diversify financial risks, financial derivatives have emerged and

developed to a large extend, especially when they are used as a tool for speculating, not hedging. As a result, global financial markets have made fundamental changes, and market risk becomes the most important risk for financial activity participants.

By definition, market risk refers to the risk of losses in institutions' (i.e., banks') trading book due to the existence of changes in equity prices, interest rates, exchange rates, commodity prices, and other indicators with values set in a public market. Market risk management is the process of circumventing, decentralizing, controlling and preventing risks by using various tools and technology on the basis of accurately identifying risk profiles and quantitatively measuring risks, among which risk measurement is a basic and essential step. The abovementioned risk measurement is to measure the amount of loss caused by quantifying the adverse changes in market factors and the characteristics of risk.

By quantifying market risk, banks and other financial institutions can recognize, assess and report the financial risks in their portfolio, in the sense of the level of uncertainty about future payoff. Their main purpose is the determination and insurance of those institutions with sufficient amount of reserve capital according to their portfolios in the event of extreme market condition and unexpected losses. The reserve capital, also called risk capital, is considered as a buffer with the purpose of protecting institutions from insolvency and it cannot be used unless unexpected contingencies or liquidation. At the regulatory level, supervisors such as Basel Committee on Banking Supervision (Basel Committee), also make efforts to set up standards for banking and insurance industries.

But the determinations of the "buffer", or the risk measure are challenging. On the one hand, banks and other institutions are supposed to hold appropriate risk capital in case of sudden fluctuations in the market. While on the other hand, by comparing the accumulated risks with their stated risk appetite, banks want the maximum efficiency and competitiveness out of their capital, suggesting the risk measures cannot be overly conservative and subsequently resulting in a waste of capital.

In order to set up an appropriate risk appetite, a variety of strategies have been proposed combined with financial engineering and mathematical methods. In 1995, Basel Committee stipulates Value at Risk (VaR) as a risk-modelling tool, which has then been widely used by financial industry to evaluate and predict potential losses. The advantages of VaR are that it is simple and intuitive, easy to calculate and easy to implement. However, VaR also has obvious limitations, such as it cannot deal with the extreme changes in the prices of investments, and it does not have convexity, and cannot reflect the diversification effect.

To overcome the disadvantages in VaR, Artzner (1999) introduces the measure of Conditional Value at Risk (CVaR), which is defined as the expected loss beyond VaR. The obtained CVaR is more informative compared to VaR, because it represents the average level of excess loss, that may be experienced when the loss exceeds the VaR threshold. CVaR is thus indicative of the potential loss worse than VaR. In addition, it has been shown that CVaR is a convex (Rockafellar and Uryase, 2000, 2002) and coherent (Pflug, 2000;

Acerbi and Tasche, 2002) risk measure, which also stimulates CVaR's application in practice (Topaloglou et al., 2002).

Although CVaR has performed much better than VaR in most cases, some shortcomings existing in CVaR still remain. For example, CVaR gives the same weight to all the losses beyond VaR threshold. But in reality, not only are most investors risk-averse, but their degree of subjective aversion to risk vary. To address the issue, Acerbi (2002, 2004) constructs a new type of risk measure based on CVaR: the Spectral Risk Measures (SRMs). SRMs are an integration or a sum of different risk levels according to different risk weights, representing a relatively new class of risk metrics. SRMs are more useful because they take into account the user's risk attitude and are therefore a significant improvement from CVaR. SRMs also share with the coherence the desirable property of coherent risk measures.

To calculate the risk measures, one of the requirements is having the knowledge of (future) Profit and Loss distributions. If we want to be completely accurate, then the full knowledge of distributions is needed. In practice, since it is difficult or virtually impossible to determine the actual future distribution of the scenarios, one common way to estimate the distribution is to construct the empirical distribution from the historical data. The main assumption here is that trends of past price changes will continue in the future. However, distribution estimated based on the past data may not be the one for the future and may be prone to errors. For example, distributions estimated based in 2000-2007 can be subject to

dramatic biases when compared to the realized distributions during financial crisis in 2008 and 2009. In this thesis, we address the issue by proposing a categorization and considering reasonable similar assets at the same time.

1.2 Literature Review

Risk Measures are represented as a map assigning a numerical value to each random payoff. With the Basel committee guidelines (1995), Value at Risk (VaR) has become one of the central planks in bank regulations and internal bank risk management. Even though VaR is in some sense superior to the previous risk measures (i.e., volatility), it still comes under considerable criticism from both a theoretical and a practical point of view: VaR lacks subadditivity and thus is not a coherent risk measure (Artzner et al., 1997, 1999; Delbaen, 2002). VaR has also been criticized for not quantifying the risk beyond the specific confidence level.

To overcome the difficulty (Acerbi and Tasche, 2001) of formulating the scenarios that have probability equal or greater than the level α, Artzner (1999) introduces Conditional Value at Risk (CVaR), which is proved to be a coherent risk measure (Pflug, 2000; Acerbi and Tasche, 2002) and enjoys increasing popularity in financial management (Bogentoft et al., 2001). Another breakthrough is the alternative risk measure: Spectral Risk Measures (SRMs), proposed by Acerbi (2002, 2004), which has a distinct difference from CVaR.

SRMs relate the risk measure directly to the user's risk aversion function by setting up a "spectrum".

Although VaR, CVaR and SRMs have their corresponding advantages in both theory and practice, when evaluating, they require one thing in common: the perfect knowledge of the loss distribution (Cont et al., 2010). In practice, the future payoff distribution is unknown and typically estimated from the historical data. Apparently, estimation error in the Profit and Loss distributions has a considerable impact on risk measures. By examining the sensitivity of risk measures to these errors (Gourieroux et al., 2000; Gourieroux and Liu, 2006), it appears that current risk measures, especially CVaR, are quite sensitive to observations and lack stability with respect to small changes, or even a single outlier, in the data set. Those problems motivate the need for studying robustness property of risk measures.

Robustness in risk management refers to risk measurements are stable and acceptable under small variations in the distributions (Vincke, 1999). The robust property of risk measures is important when people face the issue of distribution ambiguity, suggesting that we cannot perfectly estimate future distribution. In general, ambiguity is associated with one-to-many relations (Ellsberg, 1961), i.e., situations with two or more alternatives left unspecified [1]. So distribution ambiguity indicates that the probability distributions themselves are unknown and therefore there is potentially more than one distribution

[1] Klir, G.J., 1987. Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like? Fuzzy Sets and Systems, 24(2), pp.141.

needed to be considered. People are aware of the distribution ambiguity and address that by different ways. For example, Natarajan et al. (2009) present relationships between properly defined uncertainty within risk measures, which leads to a worst-case CVaR measure. Similarly, Bertsimas and Brown (2009) also study the relationships, but they instead focus on selecting and specifying the ambiguity based on the risk preferences of the modeler. Zhu et al. consider a mixture of distribution ambiguity (2009) and a robust CVaR relative to a benchmark (2010). An important commonality of those methods is that they all define different specific types of uncertainty, which may be inaccurate. Our method, instead, does not distinguish uncertainty types or ambiguity levels and thus avoids the potential errors.

In terms of robustness, before the widespread use of risk measures, some (Costa and Paiva, 2002; Goldfarb and Iyengar, 2003) study the parameter robustness in the Mean-Variance framework. With the rise of Value at Risk and other risk measures, literature's attention turns to the development of robustness of VaR and CVaR. There are various papers which seek to set up more robust risk measures, and existing literature can be divided into four types (Mohajerin Esfahani and Kuhn, 2018).

The first type to address distribution robustness is moment, which starts from considering all the distributions satisfying certain moment constraints. Based on first and second moments, El Ghaoui et al. (2003) and Zhu and Fukushima (2005) first investigate robust portfolio selection using Worst-case VaR (WVaR), and Worst-case Conditional

Value at Risk, respectively. The rationale behind WVaR is to analyze the worst-case impact of possible changes corresponding to different scenarios, when the distribution of returns is partially known. After WVaR is proposed, Huang et al. (2007) extend that approach to incorporate uncertain exit times in robust portfolio selection. But the problem with using moments is that the theoretically-derived distributions with the same first or second moments are often unrealized, making the above methods less practical.

Another type of robust VaR estimation is investigating the asymmetric distributions. Chen et al. (2007) introduce a new deviation measure based on asymmetry and apply it to uncertainty sets. Besides, Natarajan et al. (2008) obtain a modified, coherent VaR measure, named "Asymmetry-Robust VaR", which also incorporates asymmetries in the return distributions. Goh et al. (2012) introduce the concept of Partitioned VaR to capture asymmetry by separating P&L distributions into positive and negative half-spaces. The problem of asymmetric distributions is the same as moment approach: the asymmetry can help us set up risk measures, but hypothetical distributions may not be realized. In our thesis, we do not follow these two approaches, which only set up distributions by assumption.

The third type of work in the literature is defining confidence region of goodness-of-fit (Bertsimas, Gupta and Kallus, 2018) from a perspective of statistics, but one of its main controversies is whether the goodness-of-fit can precisely depict P&L distributions' characteristics.

This thesis is motivated by the last attractive alternative, which define a distribution set, as a multi-dimensional "ball", in the space of probability distributions by using a distance function such as the Prohorov metric (Erdo˘gan and Iyengar, 2006), the Kullback–Leibler divergence (Calafiore, 2007; Hu and Hong, 2013), or the Wasserstein distance (Pichler, 2013). Such sets contain distributions that are close to each other enough with respect to a prescribed similarity measure. By adjusting the radius of the set, we can thus control the degree of conservatism of the underlying optimization problem (Esfahani and Kuhn, 2018). And our approach continues the idea of distribution sets based on Wasserstein distance.

Wasserstein distance is chosen not only because it can measure the distance between distributions, but also due to a valuable notion of this distance: Wasserstein distance provides an upper bound for the change in the risk measure (Wozabal, 2012; Pichler, 2013).

Wasserstein distance was introduced in a general setting by Bickel and Freedman (1981) and in the finance context by Rachev et al. (2008)[2]. By using Wasserstein metric, people forecast P&L distributions (Cont et al., 2010), introduce an index of qualitative robustness (Krätschmer et al., 2011, 2012), and also provide various applications of Wasserstein distance in optimization problems.

To be more specific, Rüdiger Kiesel et al. (2016) start with qualitative robustness and convex risk measures (Krätschmer et al., 2012, 2014), and show several examples about

[2] Wasserstein distance is called L_p metric in their work.

the application of the Wasserstein metric as a core distance in stochastic models of decision-making process. Zhao and Guan (2018) apply the Wasserstein metric to construct a confidence set in a two-stage stochastic optimization framework. They also show the convergence property of their work: as the size of historical data increases, the risk-averse stochastic optimization problem converges to the traditional two-stage one. Esfahani and Kuhn (2018) construct the distribution ball centered at the uniform distribution on the training samples. Their work is focused on the computation efficiency of worst-case expectation problem. They demonstrate that problem can be solved as finite convex stochastic programs and Wasserstein set provides an upper confidence bound on the out-of-sample cost of the worst-case optimal decision. Yu and Schlog (2018) adopt the Wasserstein distance in comparing risk measures. In particular, they focus on the "transportation cost" definition of Wasserstein distance and expand the Wasserstein approach from the single-asset variance risk to the multi-asset allocation problem. But their primary concern is volatility, mean-variance matrix and Sharpe ratio, without consideration of other distribution-based risk measures.

The main differences between the thesis and abovementioned work is: purposes are different. The main goal of the abovementioned papers is either to assist the portfolio optimization, or to improve the calculation efficiency; while the main task of our work is to establish a more robust method to measure risks. Another difference is even if those

papers involve risk measures, they do not discuss VaR, CVaR or Spectral Risk Measures, which are one emphasis in our work.

When determining the distribution set, usually there is a key parameter, which is the radius that defines the size of the set. The larger the radius, the more distributions are contained; the smaller the radius, the less contained. Radius significantly impacts the degree of conservatism and accuracy of the risk measures, but meanwhile this parameter is hard to determine. The influence of the radius on robustness and the out-of-sample performance has been investigated by Pflug and Pichler (2012) and Mohajerin Esfahani and Kuhn (2018). Unlike previous works which deal with the parameter directly, we apply the clustering method to address this issue.

Clustering, conceptually, aims to categorize multiple objects into a number of homogeneous sets, or so-called "clusters" (Irpino et al., 2014), according to different similarity criteria, such that objects in the same cluster are close, and objects in different clusters are dissimilar (Subramanian and Shafer, 2001). Being simple to describe its scope, the fine results of clustering involve the clustering algorithm, the determination of the number of clusters (Barrio et al., 2019) and the dissimilarity / similarity measures for comparing data (Irpino et al., 2014).

For clustering algorithm (Jain and Subes, 1988; Jain et al., 1999), it can be divided to hierarchical clustering, which produces a nested series of partitions for splitting clusters

based on similarity, and k-means clustering, which is the simplest and most commonly used one employing a squared error criterion (McQueen 1967).

In terms of the similarity measures, the well-known distance measures include Euclidean distance, Mahalanobis distance (Mao and Jain, 1996) and Hausdorff Distance (Huttenlocher et al., 1993), etc. In this thesis, we use the Wasserstein distance as mentioned before, considering its ability to quantify distance between distributions and its close relationship with risk measures.

The Wasserstein-based Clustering can be found in the recent literature as well. Irpino and Verde (2006) and Irpino et al. (2006) argue that Wasserstein distance is an extension of Euclidean distance between quantile functions. They also show that it is possible to obtain a decomposition of the variability of a set of histograms and use such properties for extending clustering from standard data to histogram-valued data. Irpino and Verde (2006) conclude that Wasserstein distance is theoretically useful when data are represented by continuous density functions[3], and show a hierarchical clustering example performed on a climatic dataset. Later on, by studying the covariance based on Wasserstein metric, Irpino and Verde (2008) propose a Mahalanobis-Wasserstein distance for clustering histogram variables, and show its analogies with the Mahalanobis distance for standard variables.

[3] Irpino, A. & Verde, R., 2006. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. Data science and classification, IFCS 2006, pp.192.

Based on the previous work, Irpino (2018) provides a package in R, named "HistDAWass", which contains unsupervised classification techniques for distributions. It is the main analysis tool being used in this thesis as well.

The difference between those abovementioned works and this thesis is that none of them connects the Wasserstein-based clustering to risk measurement, or uses Wasserstein-based clustering to evaluate worst-case risk.

Backtesting is be used to test the validation of our method, which comprises a comparison of our result's out-of-sample risk measure forecasts and the investment's realized returns. In the last two decades, several formal backtests have been proposed in the literature (Kupiec, 1995; Christoffersen; 1998; Berkowitz, 2001). In our approach, we use the classical procedure (Basel, 2013) to test VaR, which is based on a binomial test for the number of exceedances over the VaR threshold, and a relatively new method to test CVaR and Spectral Risk Measures discussed in Costanzino and Curran (2015).

1.3 Research Question

As introduced in Section 1.2, the fact that we cannot perfectly estimate distribution leads to the issue of distribution ambiguity. In efforts to address this issue, robust formulations have been proposed. Robust risk assessment requires risk measures to be stable even with ambiguous data, meaning the final measures to be insensitive to the ambiguity. The motivation in this thesis is that since estimates from a single historical

distribution are unreliable, if we estimate from multiple similar distributions, or a distribution set, the results can be more convincing. Actually, among previous works, many authors have considered a wide range of forms for the distributional set[4], but the way that we define distribution sets are different from their works. Our idea is to follow the path of categorization, where we categorize similar assets into groups.

In terms of categorization, intuitively, people use "sector", where companies in similar industries are grouped together. This industry-based category is quite natural: companies in the same business type share the related products or services, related operating characteristics and related performance according to business competition. But unfortunately, our tests on sector classification show that this categorization is not risk-orientated: the risk measure results within the same sector are not concentrated, indicating that stocks in the same sector do not share the similar risk profile.

To evaluate risk in cases where the distribution of (future) asset returns is ambiguous and unpredictable, our thesis seeks to develop a more robust classification approach. This approach aims to estimate the extreme level of risk within distributions sets by taking into account distribution ambiguity. It also provides an estimate of the capital needed as an additional buffer to hedge the worst-case scenario. According to Wald (1950), worst-case risk measurement is a reasonable maxmin decision criterion. It is a sound approach and a good strategy: given results that only partially describe risk, one searches for the most

[4] Delage, E. & Yinyu Ye, 2010. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. Operations Research, 58(3), pp.596.

adverse scenario, or the best-possible upper bound on a risk measure[5]. Such a worst-case estimate has to be handled with great care, since if it is too large, it may include superfluous distributions and can be overly conservative; if it is too small, it does not serve the purposes, because the estimate should be rich enough to contain the true data-generating distribution with high confidence.

1.4 Contribution and Outline

In this thesis, we propose a more robust approach to measure market risk and examine how it can be applied to provide additional buffer. More specifically, to address the estimation biases from one individual historical distribution, we consider multiple similar Profit and Loss distributions at the same time. To construct a reasonable distribution set, we use clustering method, which is an effective way to categorize a set of objects based on their similarity. Moreover, we apply a novel distance function in risk management, the Wasserstein metric, which is a natural distance to compare distributions and meanwhile has a close relationship with the current risk measures. Using real-world data, we identify a Wasserstein-based clustering associated with robust frameworks. This method is especially important when the methodology can be highly sensitive to the input data and the input changes significantly from time to time.

[5] Goovaerts, Kaas & Laeven, 2011. Worst case risk measurement: Back to the future? Insurance Mathematics and Economics, 49(3), pp.381.

The thesis is organized as follows: Section 2 reviews and compares the most widely used measures for estimating market risk: Value at Risk, Conditional Value at Risk and Spectral Risk Measures, and also discuss the properties of coherent risk measure and the necessity of it. Section 3 introduces the clustering method for identifying sets in data. We briefly introduce two popular clustering algorithms, hierarchical and k-means clustering, which gives us the idea that applying clustering techniques can help estimate mixture of distributions. As an advance, in Section 4, we develop a Wasserstein-based hierarchical clustering, which employs the Wasserstein metric as the similarity distance function for distributions. Section 4 also summarizes our method framework and talks about the validation approach: backtesting for different risk measures. Section 5 applies the methodology to real-world data and uses backtesting to verify validation. Finally, Section 6 briefly concludes the thesis.

2. Risk Measures

In order to measure risk, people need to define a way how risks are going to be quantified. And the task of risk measurements has been started from the use of certain functions to evaluate risk. In this section, we first give an overview about the recently-developed risk measures, which assist us to quantify market risk from investments' Profit and Loss distributions. After briefly discussing the mathematical properties of a desirable risk measure, we review the axiomatic definition of the convex risk measure and coherent risk measure, which plays a key role in our framework. Then this section is mainly focused on the three widely-used risk measures: Value at Risk, Conditional Value at Risk and Spectral Risk Measures.

2.1 The Theory of Convex Risk Measures

Accurate measures of market risk are of great importance. In order to assess the risk level and take the proper safeguards, many methods of measuring risks have been developed. Initially, people used volatility of returns (i.e., variance, standard deviation, and mean absolute deviation) to describe the risk of fixed or known cash flows of financial assets, such as treasury bills and stocks. In 1952, Markowitz proposed the "expected return-variance of return" rule, which helps people to measure the risk of a portfolio by the joint return distribution of all assets. Later, the concepts of duration, adjustment duration and

effective duration have become common tools for measuring the risk of fixed income securities; the beta coefficient is used to measure the market risk of stock portfolios.

One of the breakthroughs made in the past decades is the definition of risk measure: the risk has been explicitly quantified by a risk measure that maps the loss to a real number. Namely, a risk measure is a function:

$$f(\,): X \to R \tag{1}$$

where X is a random loss and R is a real number.

In particular, the following axioms are proposed (Artzner, 1999) to check a proper risk measure by whether it satisfies specific desirable properties:

(**1**) **Translation invariance.** For all real number c, we have

$$f(X - c) = f(X) - c$$

(**2**) **Monotonicity.** For all $X \leq Y$, we have

$$f(X) \leq f(Y)$$

(**3**) **Subadditivity.**

$$f(X + Y) \leq f(X) + f(Y)$$

(**4**) **Positive homogeneity.** For all $\lambda \geq 0$, we have

$$f(\lambda X) = \lambda f(X)$$

where X and Y are any random losses.

The axiom of translation invariance is a natural requirement and also known as Risk-free Condition, which illustrates that adding a risk-free amount to a portfolio and investing it in the reference instrument, results in a decrease of the risk of the position by exactly the same amount. This is consistent with the common sense in the real-life: the less reserve or margin will correspond to higher risk that the banks and insurance companies will confront in the future.

Monotonicity explains that if the position Y performs always better than position X, then the risk associated to X should be higher than that related to Y. The axiom of Subadditivity reminds us of the diversification: a portfolio composed by several assets will be less or equally risky than a portfolio made up by a single instrument. This concept is very important in most financial and investment fields, and any risk measure that does not meet this axiom will encounter some problems and criticism in the industry.

Positive homogeneity is a limiting case of subadditivity, showing what happens when no diversification occurs. It also has the interpretation that the risk measure is independent of scale changes (e.g. currency changes) in the unit in which the risk is measured.

Any function $f(\ )$ in definition (1) that satisfies **(1)** and **(2)** are called monetary risk measure (Follmer and Schied, 2002), which focuses on monotonicity and cash invariance. Any monetary risk measure can be viewed as a capital requirement: the minimal capital that has to be added to the capital position to make it acceptable. One of the most popular methods - variance is not a monetary risk measure because it violates axiom **(2)**.

A risk measure which satisfies **(1)**, **(2)** and **(3)** is called a convex risk measure, indicating that diversification should not increase the risk. The assumption of convexity is justified by the reduction of the expected loss, in case of default, by diversification.

Lastly, a convex risk measure is named coherent if it is also positively homogeneous i.e., in axiom **(4)**. Due to its desirable mathematical and economic properties, coherent risk measure has received much attention in the quantitative risk management community and become a standard for introducing new risk measures since proposed by Artzner et al. (1999), who even argue that the word "coherent" is redundant and any risk measure must be coherent. However, not every prevailing risk measure satisfies those four axioms, such as Value at Risk, which is quite popular in financial industry and banking regulations, is not in line with axiom **(3)** and fails coherency.

2.2 Value at Risk

In 1993, Value at Risk (VaR) was first promoted by G30 in its report and quickly gained widespread application. The following year, with the aim to promote transparent risk management, J. P. Morgan group announced its risk metrics models, the core element of which is VaR. With the further development of financial markets, VaR methods have been widely used in the risk management of financial institutions. Since 1995, the Basel Committee on Banking Supervision (Basel Committee) has stipulated that the capital adequacy of commercial banks must include and be based on VaR.

VaR measures the possible downside loss of a financial asset or portfolio-over a certain period of time and at a certain confidence level α (typically 90% - 99%). From a mathematical standpoint, VaR is actually a quantile of the distribution of future returns or losses. Recall the definition of quantile:

$$\mathrm{q}^{\alpha}(X) = inf\{x|P(X \leq \mathrm{x}) \geq \alpha\} \tag{2}$$

Accordingly, VaR can be expressed by equation (3), where random variable X is the random loss of financial assets and α is a confidence level ($0 < \alpha < 1$):

$$VaR_{\alpha}(X) = \mathrm{q}^{\alpha}(X) = inf\{x|P(X \leq \mathrm{x}) \geq \alpha\} \tag{3}$$

The choice of confidence level α varies among different risk managers. For example, Basel Committee recommends the 99.9% confidence level, while lower confidence levels are often used for internal control (i.e., Bank Trusts disclosing 99% and Goldman Sachs using 95%). Another parameter that varies is the time horizon (as called "holding period") over which VaR is estimated. In financial industry, the length depends on the institution and always ranges from ten days to six months, some even extending to two years. Apparently, the higher the confidence level and the longer the holding period, the more conservative the VaR results, which means that the institution needs more reserve capital and can be safer theoretically.

Speaking of the estimation process, one of the most common ways is historical simulation, which simplifies the procedure and does not make any assumptions about P&L distributions. First, one needs to choose the time window of observations and confidence level α, then investment profits / losses within this window are sorted in ascending order:

$$X_{[1]} \leq X_{[2]} \leq \cdots \leq X_{[n]}$$

and the α-quantile is given by the return that leaves α of the observations on its left side and $1-\alpha$ on its right side. If such a number falls between two consecutive losses, then some interpolation rules will be applied. To compute the VaR for the following day, the whole window is moved forward by one observation and the entire procedure will be repeated.

Starting from the calculation steps, the most obvious advantages of VaR metrics are the succinct meaning of VaR and the intuitive method of estimation: it is easy to understand and can be calculated from historical data as shown. This advantage enables the market risk of the portfolio to be embodied as a figure, thus facilitating the realization of business risk management, public reporting and regulation objectives. Federal Financial Institutions Examination Council (FFIEC) in U.S. requires all the banks and similar institutions to report their previous day's VaR-based measures every day.

But VaR itself also has certain limitations: it does not satisfy the subadditivity in general and is therefore not a coherent risk measure. In fact, subadditivity is one of the most important requirements for risk measures. Its economic significance is that the

portfolio can diversify unsystematic risk: the risk of the portfolio is equal to or lower than the sum of the risks of each individual equity. That is to say, when Markowitz (1952) proposes that portfolios can be diversified and reduce investment risk, VaR, instead, tends to get opposite results. VaR of a combined portfolio can be larger than the individual portfolios, which runs counter to the market phenomenon of risk diversification, and is unreasonable in an economic sense.

Moreover, VaR only provides an estimate of loss at some confidence level and hence does not consider distribution properties beyond that certain level; the weight of VaR to any risk above the tolerant level is zero. For example, 95% VaR (i.e., 95th quantile) means that in 95% of cases the loss is expected to be smaller than the VaR amount, but it does not say anything about the size of losses in the rest 5% cases and merely provides a lower bound for potential losses, which may lead to an unintentional high risk under extreme conditions. In reality, after 2008 financial crisis, criticism of banks' VaR measures has become vociferous since many banks' reported VaR value appeared to give little forewarning of the potential losses and the high unanticipated level of realized losses (Nocera, 2009).

2.3 Conditional Value at Risk

In view of the limitations of VaR, especially its non-subadditive nature, many alternative risk measures have been proposed, among which the most widely used one is

the Conditional Value at Risk (CVaR), also named Expected Shortfall (ES), which has been proved by Acerbi and Tasche (2002) as a coherent measure and usually proposed as a supplement to VaR.

$CVaR_\alpha(X)$ is formally defined as

$$CVaR_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 VaR_\gamma(X)d\gamma \tag{4}$$

And particularly, in the case of continuous payoff of financial assets, (4) can be converted to

$$CVaR_\alpha(X) = E[X|X \geq VaR_{1-\alpha}(X)] \tag{5}$$

where $CVaR_\alpha(X)$ is equal to the conditional expectation of X subject to $X \geq VaR_\alpha(X)$.

And obviously,

$$CVaR_\alpha \geq VaR_\alpha$$

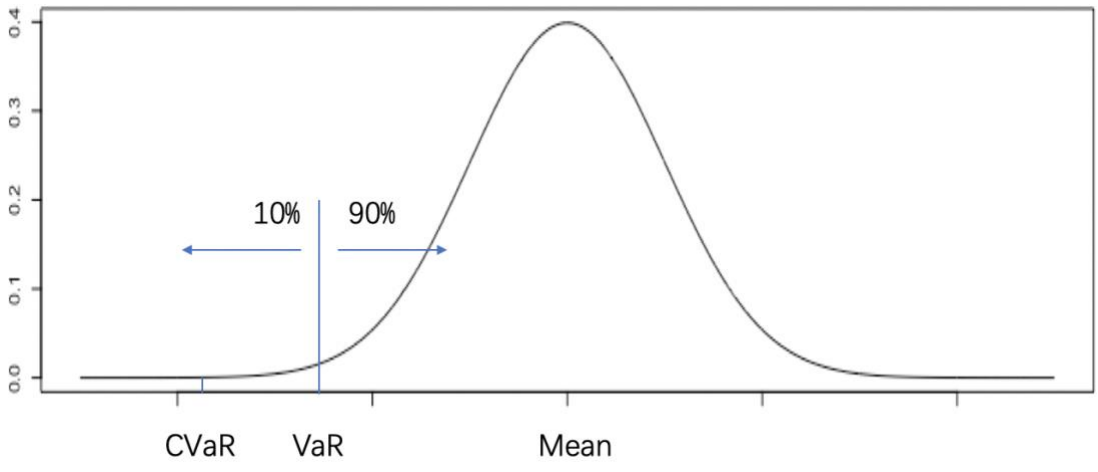


Fig. 1 90% VaR and CVaR of a hypothetical Profit & Loss[6] distribution

Applying the same historical simulation as VaR in Section 2.2, the worst $\alpha\%$ losses are then $X_{[1]} \leq X_{[2]} \leq \cdots \leq X_{[\eta]}$ and the obvious CVaR estimator is the average of the highest $\alpha\%$ losses from the set of $X_1, X_2, \ldots, X_n$:

$$CVaR_\alpha = \frac{1}{\eta}\sum_{i=1}^{\eta} X_{[i]}$$

CVaR represents the arithmetic average of excess losses and reflects the average potential loss that can be experienced when losses fall beyond the VaR threshold, providing a more conservative value and better reflecting the potential tail risk compared to VaR at the same confidence level. As shown in Fig. 2, two portfolios have the same VaR level, but a large loss is more likely in the second portfolio. CVaR can address this issue because the mean of losses above the confidence level are different, hence the two CVaRs are different.

[6] In the P&L distribution, the left tail represents the loss and the right tail represents the return. The closer the point on the curve is to the left, the greater the loss. The same below.

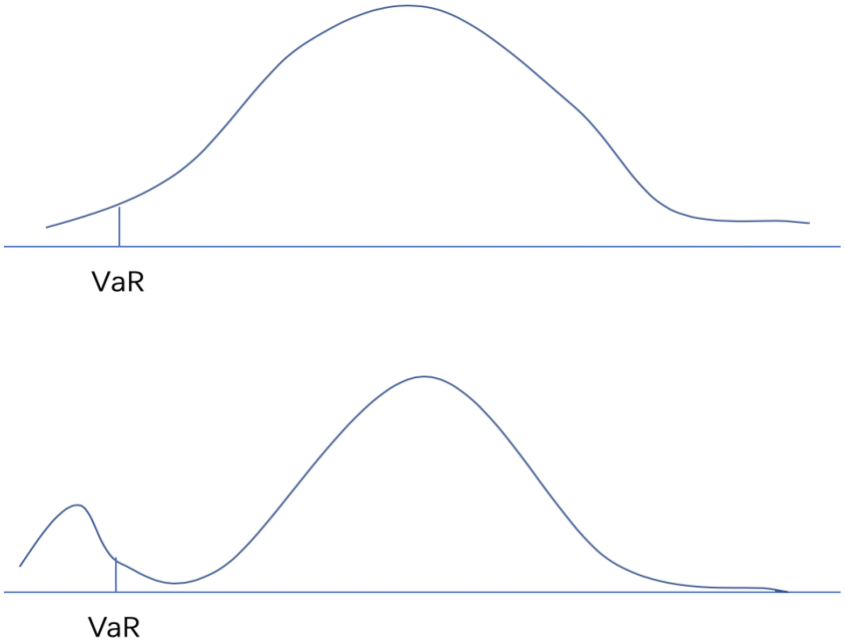


Fig. 2 P&L distributions with the same VaR but different CVaRs

More importantly, unlike VaR, CVaR satisfies subadditivity in any case-is therefore a coherent risk measure, being more reasonable in mathematical logic and financial logic.

A further motivation for CVaR is its convexity and tractability from the prospective of optimization. If there is an optimization solution at all, this solution is either a singleton or a convex polyhedron[7]. Every local optimum is global. And CVaR is a convex function of portfolio positions, allowing the construction of efficient optimizing algorithms. In particular, minimizing CVaR optimization can be decomposed and reduced to convex programming, even to solvable linear programming (Uryasev and Rockafellar, 1999; Ogryczak and Ruszczynski, 2002), which makes many large-scale calculations practical,

[7] Stanislav, P, 2000. Probabilistic Constrained Optimization: Methodology and Applications, Boston, MA: Springer US, pp.278.

efficient and stable. Basel III (2010) shifts from VaR to CVaR measure of risk, to capture tail risk, especially during market stress.

But CVaR also has some disadvantages. Under some general distributional assumptions, CVaR requires a larger sample size than VaR at the same level of accuracy (Giannopoulos and Tunaru, 2005). Meanwhile, unlike VaR, which is less sensitive to small deviations of the underlying distribution, CVaR is more sensitive to the tail of the distribution and estimation errors. Accuracy and robustness of CVaR estimation heavily relies on accuracy of the tail modelling (Kou, Peng and Heyde, 2013) and sample size (Cont et al., 2010). Thus, the risk arises due to the uncertainty of the underlying P&L distributions, which can readily be observed in the case where enough data samples are not available, or the data samples are unstable[8]. Since under extreme market conditions, the stable relationship between various financial factors has been destroyed, CVaR estimates given at that time may have large deviations.

Another limitation of CVaR is that it gives an equal weight to all the risks above the designated level, which is considered over-simplified. In reality, people are not only risk averse, but also give different (higher) weights to higher degrees of risk, so it is natural to work with a risk measure that takes account of risk aversion level. This takes us to the class of Spectral Risk Measures (SRMs) of the underlying probability distribution as a coherent generalization to CVaR.

[8] Shushang Zhu & Fukushima, M, 2009. Worst-Case Conditional Value-at-Risk with Application to Robust Portfolio Management. Operations Research, 57(5), pp.1156.

2.4 Spectral Risk Measures

To address the shortcomings of CVaR and relate the risk measure to the user's risk aversion, Spectral Risk Measures (SRMs) were proposed by generating a weighted average of the quantiles of a loss distribution, the weights of which depend on the user's risk aversion, as stated by Acerbi (2002, 2004):

$$\mathrm{SRM}(X)_{\phi} = \int_0^1 VaR_{\gamma}(X)\phi(\gamma)d\gamma \tag{6}$$

where $\phi(\gamma)$ is the weight, the density function given to different possibility γ, known as the "spectrum".

Especially, VaR places all its weight on the α quantile and gives every other outcome a weight of zero; CVaR gives all tail quantiles the same weight of $\frac{1}{1-\alpha}$ and gives non-tail quantiles a weight of zero. In other words, CVaR is a special case of Spectral Risk Measures where

$$\phi(\gamma) = \begin{cases} \dfrac{1}{1-\alpha}, & if\ \gamma > \alpha \\ 0, & if\ \gamma \leq \alpha \end{cases}$$

for a certain probability $\alpha \in (0,1)$, meaning that $\phi(\gamma)$ is simply uniform in $\gamma \in [0, \alpha]$ and zero elsewhere.

To conclude, the VaR is based on a degenerate weighing function and the CVaR is based on a simple step weighting function. Neither of those two risk measures makes any

allowance for the user being risk-averse (Grootveld and Hallerbach, 2004). But SRMs provide a more flexible approach to reweight the quantiles.

In a more general case, the function $\phi(\gamma)$ assigns different weights ϕ to different "γ-confidence level slices" of the left tail of the P&L distribution, even to the whole distribution, including the profit part. In equation (6), the interval for definite integral is 0 to 1, which indicates that one can assign ϕ to all the returns (gains or losses), from the possibility 0.01 to 0.99. Therefore, whereas VaR and CVaR only consider the downside risk (tail part) of the P&L distribution, Spectral Risk Measures is more comprehensive by including all the possible payoffs according to one's risk preferences.

The important parameter $\phi(\gamma)$ reflects the user's risk aversion. More precisely, following Acerbi (2004), we can define the subset of $\phi(\gamma)$ satisfying the following properties (Acerbi, 2002, 2004) of nonnegativity, normalization and increasingness:

Nonnegativity: $\phi(\gamma) > 0$.

Normalization: $\int_0^1 \phi(\gamma)\mathrm{d}\,\gamma = 1$.

Increasingness: $\phi'(\gamma) \geq 0$.

The first condition requires that the spectrums are nonnegative. The second requires that the probability-weighted weights should sum to 1 and is also for the translational invariance condition. The third requirement is the most important one: the weights attached to higher losses should be no higher than or equal to the weights attached to lower losses, and is intended to reflect risk-aversion.

A drawback with increasingness property is that it does not rule out risk-neutral situations. To rule out such cases, increasingness can be replaced with the slightly stronger condition:

Strict Increasingness: $\phi'(\gamma) > 0$[9].

which ensures that the spectrum $\phi(\gamma)$ rises with γ. In 'well-behaved' cases, $\phi(\gamma)$ will rise smoothly for users who are less risk-averse and rise sharply for users who are more risk-averse.

Compared to CVaR, which is over-simplified, Spectral Risk Measures take into account the user's (i.e., the bank's or the clearinghouse's) attitude towards risk by using weighted average of the quantiles. It can be agreed that if a user is more risk averse, other things being equal, then that user should face a higher risk, as given by the value of the SRMs[10].

Besides, Spectral Risk Measures satisfy monotonicity, positive homogeneity and translation invariance, and also other properties, which are law-invariance and comonotonicity (Denneberg, 1994; Kusuoka, 2001) as defined below:

Law-invariance: For x and y with cumulative distribution functions F_x and F_y, if $F_x = F_y$, then f(x) = f(y).

[9] Dowd, K., Cotter, J. & Sorwar, G., 2008. Spectral Risk Measures: Properties and Limitations. Journal of Financial Services Research, 34(1), pp.63.

[10] Dowd, K., Cotter, J. & Sorwar, G., 2008. Spectral Risk Measures: Properties and Limitations. Journal of Financial Services Research, 34(1), pp.61.

Law invariance is important in practice as it is required for the estimation of the risk measure from empirical data. Coherent risk measures that are not law invariant cannot be estimated solely from data. Using such measures can lead to different risk values for two portfolios with identical loss distributions[11].

Comonotonicity: $f(x + y) = f(x) + f(y)$ for every comonotonic random variables x and y.

And also, Spectral Risk Measures share the property of subadditivity and thus belong to the family of coherent risk measures. SRMs therefore have those highly attractive properties of such measures as discussed earlier.

Those above-mentioned characteristics make Spectral Risk Measures more and more popular and SRMs can be applied to many different problems. Studies have suggested using them to set capital allocation requirements (Overbeck, 2004), to obtain optimal risk-return tradeoffs (Acerbi, 2004), and to set margin requirements for futures clearinghouses (Cotter and Dowd, 2006).

The thing that remains a question here is how to specify the weight $\phi(\gamma)$. The most natural way (Bertsimas et al., 2004) to obtain $\phi(\gamma)$ is from the user's utility function, and a popular choice is the exponential utility function:

$$U(x) = -e^{-kx} \tag{7}$$

[11] Balbás, A., Garrido, J. & Mayoral, S., 2009. Properties of Distortion Risk Measures. Methodology and Computing in Applied Probability, 11(3), pp.388.

where $k > 0$, also called the coefficient of Absolute Risk Aversion (Saha, 1993).

Then the weighting function can be set as

$$\phi(\gamma) = \lambda \mathrm{e}^{-\mathrm{k}(1-\gamma)} \tag{8}$$

where λ is a positive constant. An exponential risk aversion function is non-negative for any quantile, and it is increasing as moving towards a higher loss quantile. Then by integrating $\phi(\gamma)$ from 0 to 1 and setting the integral to 1 (see the Normalization property), λ can be solved as

$$\lambda = \frac{k}{1 - \mathrm{e}^{-\mathrm{k}}} \tag{9}$$

And substituting equation (9) into (8) gives the exponential weighting function:

$$\phi(\gamma) = \frac{k\mathrm{e}^{-\mathrm{k}(1-\gamma)}}{1 - \mathrm{e}^{-\mathrm{k}}} \tag{10}$$

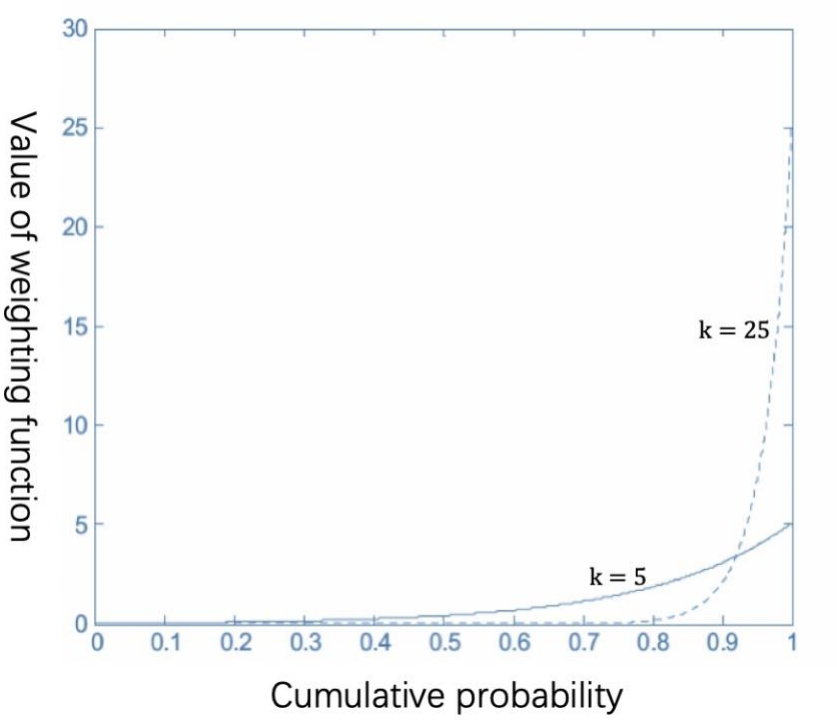


Fig. 3 Exponential weighting functions $\phi(\gamma)$ when absolute risk aversion $k = 5$ and $k = 25$, plotted against the cumulative probability γ

This weighting function is illustrated in Fig. 3 for two alternative values of k. Observe that this weighting function has a nice shape and rises exponentially with γ. In addition, for the higher γ values associated with higher losses, the weights are higher and the rate of increase of $\phi(\gamma)$ is higher, the greater the value of k. This is consistent with one common sense in our risk culture: the coefficient of relative risk aversion increases with wealth. And although we can assign weight to the whole distribution, including gains and losses, people are more disgusted with higher losses, especially the extreme situations.

SRMs can be obtained from other utility functions, such as the power utility function (Dowd et al., 2008). But till to date there is little guidance on the choice of the weighting

function. The general lesson is that SRMs users must be careful to choose the weight (or utility function) that suits the characteristics of particular problems that people work on.

3. Clustering

The clustering, especially the hierarchical clustering, is the main body of our approach, which gives us a way to consider similar investments at the same time and obtain the worst-case risk within one category. This section discusses the categorization (hierarchical clustering) technique that we use, including the similarity measure within categories and the way to generate categories-clustering algorithm.

3.1 Overview

Broadly speaking, clustering is used for organizing objects into groups whose members are similar in some way. Sector is a naïve approach to cluster investments according to their industry, but does not work well from a risk management prospective.

In finance, there are many ways to classify assets. Some (Farrell, 1974; Martin and Klemkosky, 1976) refer to the dependence on the business cycle and the state of the market, based on some priori grounds or chosen financial indicators. However, those classical methods do not guarantee the groups are homogenous according to their risk-return profiles (Vermorken et al., 2010; Costa and de Angelis, 2011). Some use wavelet analysis (Fernandez, 2005, 2008), which also relates to the defensiveness to business cycle, but only focuses on scale-dependent systematic risk.

Some may start with examining beta coefficient (Koesterich and Morillo, 2013) or variance (Conover et al., 2008; Backus et al., 2010) by using a statistical assessment of the

dependence between equites. The problem here is that either the measures (beta or variance) themselves are exposed to somehow different risk factors, or we need to assume a coverage of these factors (Bruzda, 2017).

Considering fluctuations in asset prices, some also categorize by price changes, such as defining price bubble types (Kubicová and Komárek, 2011), utilizing panel model (Adalid and Detken, 2007) or setting up regression functions for price change and other possible factors (Rae and Noord, 2006).

Another approach to categorize assets regards profit or loss data as time series. Motivated by the statistical method, some improve Pearson's correlation coefficient for time series. Some use neural networks, i.e., self-organizing maps (Kohonen, 1990), but it does not work well with time series of unequal length. Some use Dynamic Time Warping (DTW) for comparing discrete sequences of continuous values, but it turns that its result is poor for indicating unknown utterance or time-inhomogeneity because it requires evaluation of a given recurrence function.

Technically speaking, the abovementioned work about categorization can be viewed as a different way of clustering in some sense. Conceptually, clustering is a fundamental unsupervised learning method to divide a large group of observations into smaller subsets based on their similarity measured with a distance function. Compared with other techniques which are used to divide data objects into groups, its main distinct property is

that clustering derives the classification only from the data; for this reason, clustering belongs to an "unsupervised" method, in contrast to some supervised classification, i.e., objects are assigned using the model developed from the previous results.

Most clustering analysis is undertaken with the objective of addressing the heterogeneity of the data. Rather than deal with one large group of widely divergent objects, people explicitly divide the group into more homogeneous groups so that objects in the same group show a high similarity whereas objects in different groups are typically more dissimilar, where such groups are called "clusters".

There are several parameters to tune for successful clustering, in particular, the number of clusters, the clustering algorithm and the metric or the similarity structure. Some main similarity measures will be introduced in the following section. We discuss several widely-used approaches of clustering in Section 3.3, with the emphasis on two different types: hierarchical methods and k-means methods.

3.2 Distance Measures

The choice of distance measures is a critical step, and also a big challenge in clustering, which will determine how similarity of two elements is calculated and the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another[12]. To serve the purposes, people would expect that

[12] Pandit, S. & Gupta, S, 2011. A Comparative Study on Distance Measuring Approaches for Clustering. International Journal of Research in Computer Science, 2011(25), pp.29.

distance between objects within a cluster should be minimum and distance between objects within different clusters should be maximum.

In most existing data clustering method, the similarities are measured in one of the following distances:

Distance Name	Definition
Euclidean distance	$\sqrt{\sum_{k=1}^{n}(X_{i,k}-X_{j,k})^2}$
Manhattan distance	$\sum_{k=1}^{n}\left\|X_{i,k}-X_{j,k}\right\|$
Chebyshev distance	$\max\left\|X_{i,k}-X_{j,k}\right\|$
Minkowski distance	$\sqrt[p]{\sum_{k=1}^{n}\left\|X_{i,k}-X_{j,k}\right\|^p}$
Hamming distance	$(\#\frac{X_{i,k}\neq X_{j,k}}{n})$

Table. 1 Distance function in clustering

Euclidean distance is the most frequently-used one, especially applied to standardized data. It also assigns an equal weight to each standardized variable, so that after standardization, it can determine the closeness when every variable is equally important.

Manhattan distance is based on absolute value distance, as opposed to squared error (Euclidean) distance. In practice, the clustering results from Euclidean and Manhattan distances tend to be similar, but Manhattan distance shows better performance in terms of

less computation time (Jain and Dubes, 1988). One disadvantage is that it depends upon the rotation of the coordinate system (Ajiboye and Olufadi, 2018). And if Manhattan distance is chosen, the effect of single large differences (outliers) is dampened, since they are not squared.

Similarly, Chebyshev distance calculates the maximum of the absolute differences. It is still computationally efficient (Potolea et al., 2011), but tends to have more iterations (Tendolkar et al., 2015). It can be used when two objects are different on any one of the dimensions.

Minkowski metric is a generalization of the above-mentioned distances. People can manipulate the value of p to control the progressive weight that is placed on differences on individual dimensions, i.e., Manhattan distance ($p = 1$), Euclidean distance ($p = 2$) or Chebyshev distance ($p = \infty$). All Minkowski metrics are based on the location of data in such a "space", compared to other distances which based on properties, not location of data.

Hamming distance is the one between two strings with the same length and is the number of positions at which the corresponding symbols are different. In another way, it measures the minimum number of substitutions required to change one string into the other. Therefore, Hamming distance is widely used to identify groups for text or other non-numeric data, when datasets are binary in nature (Jain and Dubes, 1988).

The choice of distance function has been contested in cluster analysis, and there is no definitive standard yet. Each of these distances would favor some types of clustering as we

discuss. When determining it, one must consider the objective, the amount of data and on the complexity of the task, etc.

Among all the determinants, there is one key point worth mentioning: the structure and type of data being analyzed. For all the numerical data points, we can choose from Minkowski distances and then further determine the value of p. For text data in telecommunication, the Hamming distance can be chosen. For vector, Chebyshev distance, often defined on a vector space, can be used, where the distance between two vectors is the greatest of their differences along any coordinate dimension. Besides, for one special data type: time series, correlation coefficient is applied.

In our method, we need a distance function that can be used to compare Profit & Loss distributions, but none of the above distance formulas can describe the distance between distributions. To address this issue, we introduce in Section 4 a novel distance function: Wasserstein metric, which can be particularly helpful to quantify similarity among distributions, and even has some other desirable characteristics in terms of risk management.

3.3 Clustering Algorithms

As there are various clustering algorithms, this section introduces a framework of the clustering algorithms found in the literature into distinct categories. When determining the clustering algorithm in this thesis, we choose from hierarchical and centroid-based (k-

means) methods, which are two main types of clustering algorithms. And we finally use hierarchical clustering due to the reasons stated in Section 3.3.2.

Here is a brief introduction to those two algorithms:

(1) Hierarchical clustering:

The result of hierarchical clustering will be represented as a hierarchical tree structure, where the k-cluster solution is formed by joining together two clusters from the $k+1$ cluster solution. The idea is that objects are more related to nearby objects than those further away. The process continues until a stopping criterion is reached (frequently, the requested number k of clusters). The hierarchical method has a drawback though, which relates to the fact that once a step (merge or split) is performed, this cannot be undone[13].

(2) Centroid-based clustering:

This algorithm model clusters by a central vector, which does not need to be an actual object, and users need to define the number of clusters in advance. It is often used, primarily due to popularity and widespread use of k-means clustering (Hartigan, 1975), which forms the basis for centroid-based clustering techniques, and is conceptually easy to understand.

There are also other clustering algorithm types such as density clustering, with the aim of identifying areas of higher density than what can be found in the remainder of the data space. Other methods have been developed, which can lead to a k-cluster solution with overlapping cluster (where one object is assigned to more than one cluster and where the

[13] Fahad, A. et al., 2014. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Transactions on Emerging Topics in Computing, 2(3), pp.268.

clusters are not necessarily nested to form a hierarchical tree) and fuzzy clusters (where the assignment of an object to a cluster is a number between 0 and 1). As we only need the simple structure of clusters and the type of our clustering objectives, those methods are not considered or elaborated here.

3.3.1 Hierarchical Clustering

Hierarchical clustering models are among the earliest techniques developed. In general, hierarchical methods approach the data in one of two ways: bottom-up (called agglomerative methods, beginning with each observation in a separate cluster and joining clusters together at each step of the process until only one cluster of size n remains) and top-down (called divisive methods, beginning with all observations in a single cluster and dividing one cluster into two at each step of the process until n clusters of size 1 remain). Some methods are neither agglomerative nor divisive, for example, some using least squares to fit certain tree structures.

The main difference between agglomerative method and divisive method is that the former uses the distance function to fuse nearby clusters, while the latter uses it to split insufficiently coherent clusters. It leads to the major difficulty when using divisive clustering: knowing where to split. There are $(2^{n-1}-1)$ ways to split n objects into two subsets (Edwards and Cavalli-Sforza, 1965), so it is very time-consuming to establish a splitting protocol of all possible bipartitions. In addition, there may be some cross-over or

reversal of the branches (Roux, 2018) in divisive tree result, making it even harder to identify clusters.

As a result, the clustering method that is used in this thesis, as well as the following discussion of hierarchical clustering in this subsection, are mainly focused on agglomerative clustering.

The basic idea behind agglomerative clustering is simple. Starting with each object in its own separate cluster (i.e., n clusters of size 1), at each stage of the process, find the two "closest" clusters and join them together. Continue until one cluster of size n remains. The algorithm is simple and remarkably efficient. The iterative process is as followed:

Step 0. Start with all objects in separate clusters (i.e., n clusters with one object in each). Denote these clusters $C_1, C_2, C_3, \dots, C_n$. In this initial step, the distance between two clusters is defined to be distance the two objects they contain; that is

$$d_{C_i,C_j} = d_{i,j}$$

Let $\mathrm{t} = 1$ be an index of the iterative process.

Step 1. Find the smallest distance between any two clusters. Denote these two clusters C_i and C_j.

Step 2. Amalgamate clusters C_i and C_j to form a new cluster denoted C_{n+t}.

Step 3. Define the distance between the new cluster C_{n+t} and all remaining clusters C_k as follows:

$$d_{C_{n+t},C_k} = min\{d_{C_i,C_k}, d_{\mathrm{j},C_k}\}$$

where d_{C_i,C_k} (d_{j,C_k}) is the distance between the new cluster and each of the old cluster.

Step 4. Add cluster C_{n+t} as a new cluster and remove clusters C_i and C_j. Let $\mathrm{t} = t + 1$.

Step 5. Return to Step 1 and continue until one cluster remains.

Ultimately, all applications of agglomerative clustering end up at the same point: one large cluster consisting of all observations. In fact, it is not the endpoint of the analysis that is particularly useful, but the sequence of steps that describes which objects are joined together at what stage of the analysis (Sneath and Sokal, 1973). The graphical depiction of these steps is known as a dendrogram, which corresponds to a hierarchical tree structure generated by the iterative sequence described above.

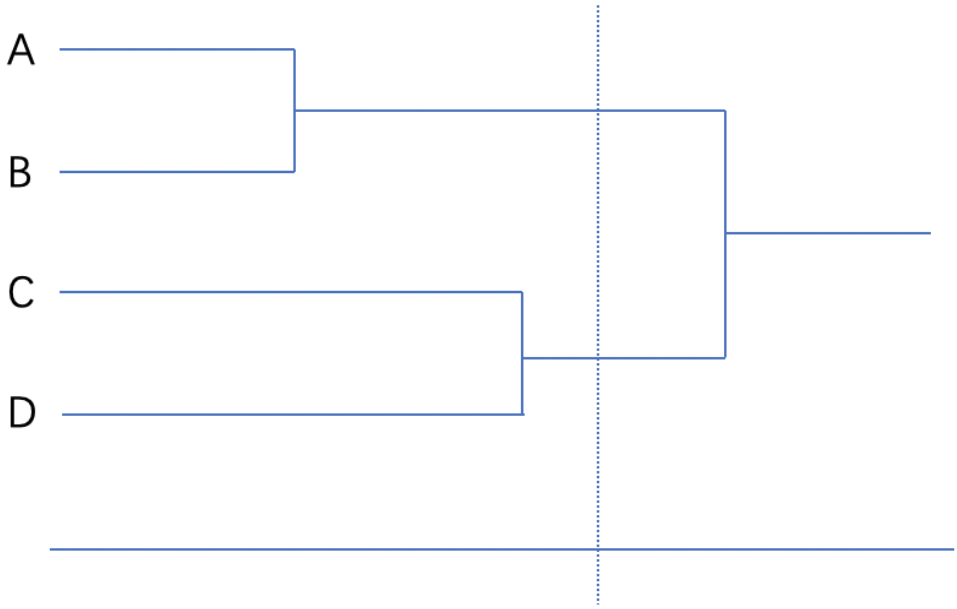


Fig. 4 Dendrogram for clustering solution

As shown in Fig. 4, at the beginning, each object A - D is considered as a cluster C, also called a "singleton". Then at each iteration, a new cluster is created by merging two closet objects or two closet clusters, which can be depicted by a vertical line connecting the two horizontal lines at the distance where the two are clustered together.

Recall that one key point in clustering is the number of clusters. In agglomerative clustering, the result does not provide a definitive answer to the question. Actually, the dendrogram is a graphical representation of a hierarchy of nested cluster solutions: a one-cluster solution, a two-cluster solution, etc., all the way up to an n-cluster solution. Drawing a vertical line on the dendrogram, corresponding to a particular distance value d, reveals the cluster result at the level of distance and the membership of the different cluster. But only looking at the dendrogram, we cannot see the "best" number of clusters. Reading and determining the cluster number from dendrogram still involves a considerable amount of subjectivity and requires great judgment on the purposes of analysis.

The above-mentioned clustering algorithm is also computationally efficient (Murtagh and Contreras, 2012): as the number of objects n increases, the worst-case amount of computational effort required increases on the order of n^2. The algorithm is even more efficient for sparse data (Matthew, 2007; Krzakala Florent, et al., 2013), such as network structure, where each object is connected to only a fraction of the other objects in the set.

One drawback of this clustering is that it tends to be extremely myopic. An object will be added to a cluster so long as it is clustered to any one of the other objects in the cluster,

even if it is relatively far from all the others (Pruscha, 2006). Thus, the algorithm tends to produce long, stringy clusters and nonconvex cluster shapes. If the true underlying clusters are nonconvex, then this property is not that important; however, in most cases the data will be convex and compact. As a direct result, the approach does not perform well in Monte Carlo studies (Milligan, 1980).

3.3.2 K-means Clustering

Centroid-based clustering is different from agglomerative clustering. Here the goal is to divide the samples into a predetermined number k of nonoverlapping groups so that the objects within each group are relatively similar and the objects between groups are relatively dissimilar. As introduced before, the most popular approach is k-means clustering (Hartigan, 1975). This algorithm is also simple to understand and efficient to compute. The steps of the general k-means clustering algorithm are described below:

Step 0. Set an initial partition of the data into k clusters.

Step 1. Calculate the centroid for each cluster C, $\bar{X}_C$.

Step 2. Each object is assigned to its nearest centroid, based on the distance calculated between those two.

Step 3. Re-compute the centroid by taking the mean or sum of square of all objects assigned to that centroid's cluster.

Step 4. Iterates between Step 1 - 3 until a stopping criterion is met (i.e., no objects change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

Those 5 steps can be summarized to two main iterative steps: cluster assignment and centroid step movement.

The main advantage of k-means clustering is that it is easy to implement, with the results presented in an easy and simple manner. Especially for big data, it is efficient in processing a large number of datasets. And unlike hierarchical clustering, k-means can easily adjust to the changes in dataset. If any problems occur, adjusting the cluster segment allows changes to easily applied to the algorithm.

However, what is more worth mentioning is k-means' shortcomings. Firstly, k-means clustering is prone to finding only locally optimal solutions because it is based on a heuristic that makes only local improvements to a starting partition until no further improvements are possible. This feature leads to a major flaw in the practical application of k-means: with different starting points, k-means clustering gives varying results on different runs. The random choice of cluster "seeds" has a strong impact on the quality of results and yields different final clusters, resulting in inconsistency. It is therefore important to give some consideration to the choice of the initial cluster partition. In general, the better the starting point, the better the final solution. This turns out to be even more

important when the data set is relatively small. In practice, people may need to re-run k-means a large number of times with different starting points to ensure a good solution.

Secondly, k-means is sensitive to the order of data (the way in which data is ordered in building the algorithm) and the scale of data. Rescaling data through normalization or standardization changes results completely. Compared with hierarchical clustering, k-means is quite unstable.

Lastly, k-means algorithm finds a cluster solution for given value of k, so it is up to the user to decide what value of k yields a better solution. Although there are some existing criteria to help people specify k, such as the sum of squared errors (SSE) (Andrew, 2012), it is still difficult to predict the number of cluster number. The final decision always involves some trade-off between the simplicity of the solution, where a smaller number of clusters is better, and its adequacy, where more clusters is better when one wants to reduce within-group heterogeneity. Just like the staring points, users usually need to conduct analyses for several different values of k and choose the solution that best corresponds to their objectives. The whole process, as we discuss, may be quite time-consuming and inefficient.

In our method, we use hierarchical (agglomerative) clustering after considering both hierarchical and k-means clustering. There are two main reasons:

(1) Hierarchical clustering can be applied to various data type, not limited to numerical data as k-means. Moreover, the dendrogram structure can give us a clearer and more comprehensive overview of the cluster structure, as well as the distance between different clusters.

(2) As discussed before, it is difficult to determine the starting "seeds" in k-means clustering, and we do not want our results inconsistent. So here we choose hierarchical clustering, which does not have the issue with unstable results.

4. Methodology and Validation Approach

4.1 Framework

Broadly speaking, instead of estimating risk based on a single distribution, which is likely biased, our method here considers multiple distributions at once, representing other possible distributions (scenarios) that can be realized in the future. The key of our method is to properly characterize these distributions so that they can be well justified as reasonable scenarios needed further consideration. In particular, while these distributions are not the same as the historical distribution, they should be "similar" to the historical distribution to some extent so that they represent reasonable or meaningful variants of the historical distribution. To achieve this, we apply the measure of Wasserstein distance to define the similarity. Then we apply clustering to identify a number of distribution sets, where each set consists of distributions that are most similar to each other. Finally, we propose to evaluate the worst-case risk by evaluating the largest possible risk resulting from all distributions in the same cluster.

4.1.1 Wasserstein Metric and Its Connection to Risk Measures

All the risk measures in Section 2 reflect the risk attribute of a single P&L distribution, which may be biased. To address the problem of distribution ambiguity, we introduce in this section the use of Wasserstein distance to determine the set of distributions, in which we can have more similar distributions in terms of their risk profiles. Wasserstein distance

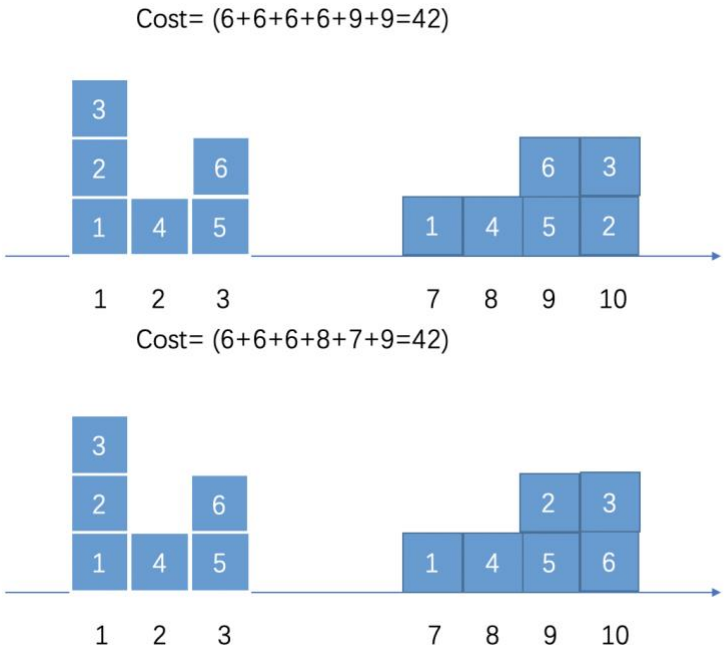
is known in the computer science community as the Earth Mover's Distance, and has been successfully used in a variety of applications including image processing (Rubner et al., 2000), shape recognition (Gangbo and Mccann, 2000) and even thermodynamics (Carrillo et al., 2006).

In the field of risk management, as introduced in Section 1.2, some use Wasserstein metric to define "qualitative robustness" (Krätschmer et al.,2012, 2014) and extend the concept to different areas in finance (Rüdiger Kiesel et al., 2016). Yu and Schlog (2018) also use Wasserstein distance to compare risk measures, with the main focus on volatility and mean-variance. Wasserstein distance is also used in optimization, in the framework of finite convex stochastic programs (Esfahani and Kuhn, 2018) and two-stage stochastic optimization (Zhao and Guan, 2018).

The rth Wasserstein distance, d_r, is defined as the rth root of the total cost incurred when transporting a pile of mass into another pile of mass in an optimal way, where the cost of transporting a unit of mass from μ_1 to μ_2 is given as the rth power $\|\mu_1 - \mu_2\|^r$ of the Euclidean distance.

For example, in Fig. 5, if we get 6 boxes and we want to move them from the left to the locations marked by the dotted square on the right. For box 1, we move it from location 1 to location 7, and the moving cost equals to its weight times the distance. For simplicity, we will set the weight to be 1. Therefore, the cost to move box 1 equals to 6. Here we have two different moving plans π to illustrate how boxed are moved. To be more specific, in

π_1, we move 2 boxes from location 3 to location 9 and the entry $\pi(3,9)$ is set to two. The total transport cost of either plan below is 42. However, not all transport plans bear the same cost, and the Wasserstein distance is the cost of the cheapest transport plan.



π_1

	7	8	9	10
1	1	0	0	2
2	0	1	0	0
3	0	0	2	0

π_2

	7	8	9	10
1	1	0	1	1
2	0	1	0	0
3	0	0	1	1

Fig. 5 Wasserstein distance between two box masses

In the distribution world, Wasserstein distance answers the geometric "assignment question" (Carlsson et al., 2018): how we can transport mass with one distribution to have another distribution (see Fig. 6), with minimal global transportation cost. In other words, it is the minimum cost of transporting mass in converting one data distribution to another data distribution. In general, this distance creates a natural way to compare distributions.

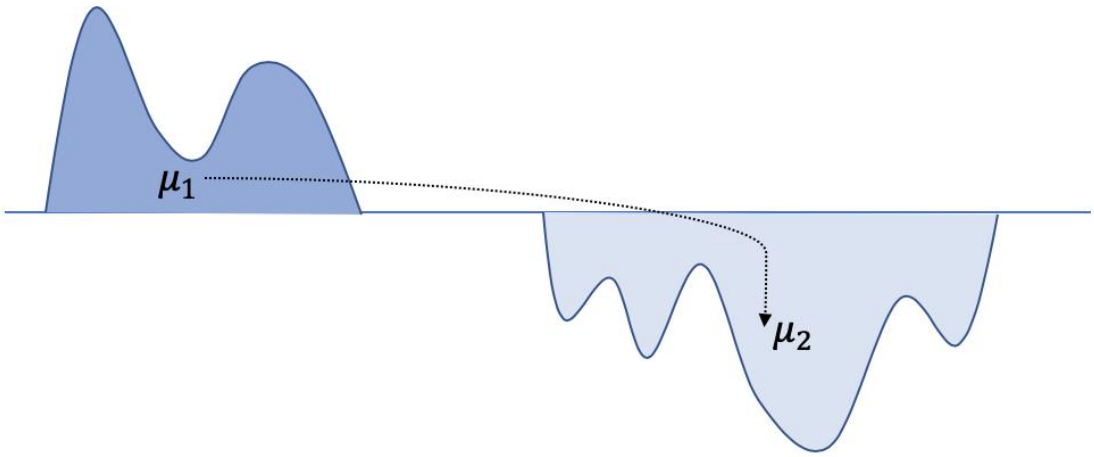


Fig. 6 Wasserstein distance between two univariate distributions μ_1 and μ_2

Another advantage of Wasserstein distance is that we don't need assumption on the form of the distributions, and distributions are observed through samples. Wasserstein distance can be estimated by solving a discrete version of equation (11) which is a linear programming problem.

Technically speaking, the Wasserstein distance between two objects can be mathematically defined as follows: let (X, d) be a Polish space (complete and separable metric space) and P, P' be two distribution functions, rth Wasserstein distance, $d_r(P, P')$ for P, P', is given by

$$d_r(P, P') = (\inf_{\pi} \int_{X \times X} d(x, x')^r \, \pi(dx, dx'))^{\frac{1}{r}}, r \geq 1$$

where the infimum (inf) is taken over all probability measures π, which is the set of all joint measures on $X \times X$ with marginals P and P', that is

$$\pi(A \times X) = P(A)$$

and

$$\pi(X \times B) = P'(B)$$

for all measurable sets $A \subset X$ and $B \subset X$. Recall Fig. 5, π contains all the possible transport plans π_i, and the $\pi(3,9)$ means how many boxes at location 9 is from location 3. The number of boxes in location 9 must originally come from any position, i.e., $\sum \pi(*,9) = 2$, which is the same as saying $\pi(\mathrm{a},\mathrm{b})$ must have marginals P and P' respectively.

In our thesis, the research object is Profit and Loss distribution for different investments, which is a univariate sample. Given two vectors a and b, their Wasserstein distance of order p can be simplified and computed from their empirical distributions (Rüschendorf, 2001). Recall the concept of "transportation cost", the vector a represents the locations on the real line of m deposits of mass $\frac{1}{m}$ and the vector a the locations of n deposits of mass $\frac{1}{n}$. Users can also specify the vectors of weights for a and b, the same as Fig. 5, where we set the weight to 1. In terms of the empirical distribution function $\mathrm{F(t)} = \sum_{i=1}^{m} w_i^{(a)}\, 1\{a_i \le t\}$ of locations a_i with normalized weights $w_i^{(a)}$, and the corresponding function $\mathrm{G(t)} = \sum_{j=1}^{n} w_j^{(b)}\, 1\{b_j \le t\}$ for b, the Wasserstein distance is given as

$$d_r(F,\mathrm{G}) = \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^r \,\mathrm{du}\right)^{1/r} \tag{11}$$

where F and G are generalized cumulative distribution functions of a and b, F^{-1} and G^{-1} are generalized inverses and $r \in [1, +\infty)$ (Schuhmacher, 2019). Note when $r = 1$, we have a particularly simple dual representation:

$$d_1(F, \mathrm{G}) = \int_{-\infty}^{\infty} |F(x) - G(x)| dx$$

the resulting distance belongs to the family of integral probability metrics (Sriperumbudur, 2010).

Another widely-used order is 2:

$$d_2(F, \mathrm{G}) = \sqrt{(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^2 \, \mathrm{du})}$$

For example, here we assume two uniform distributions as F and G respectively:

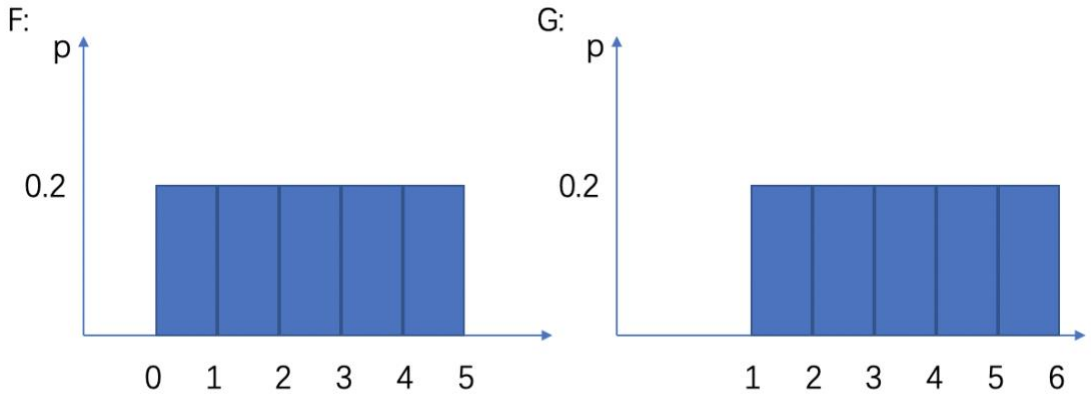


Fig. 7 Wasserstein distance ($r = 2$): example

According to the definition of Wasserstein distance, we then have

$$d_2(F, \mathrm{G}) = \sqrt{\left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^2 \,\mathrm{du}\right)}$$

$$= \sqrt{\int_0^{0.2} |1-2|^2 \,\mathrm{du} + \int_{0.2}^{0.4} |2-3|^2 \,\mathrm{du} + \int_{0.4}^{0.6} |3-4|^2 \,\mathrm{du} + \int_{0.6}^{0.8} |4-5|^2 \,\mathrm{du} + \int_{0.8}^{1} |5-6|^2 \,\mathrm{du}}$$

$$= 1$$

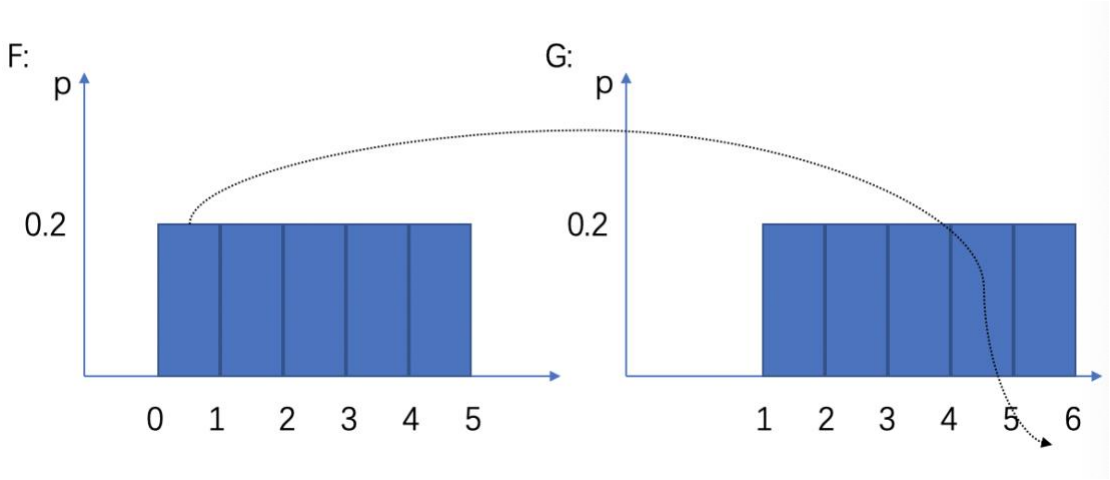


Fig. 8 Wasserstein distance ($r = 2$) and transportation cost: example

Recall the previous discussion about transportation cost, the Wasserstein distance between those two distributions can also be interpreted as: we move the first "box" from $[0,1)$ to $[5,6)$, with distance 5 and weight 0.2, and the transportation cost is also $5 \times 0.2 = 1$.

Speaking of risk measures, it turns out that the Wasserstein metric is a useful notion of distance which ensures the desired properties (Pichler, 2013): many risk measures allow an estimation in terms of the Wasserstein distance, and important risk measures are

continuous with respect to this distance. To be more specific, Wasserstein distance provides a valuable concept in the evaluation of risk functionals, which gives precise and useful bounds:

$$\left|\rho_{\phi,P}(X) - \rho_{\phi,P'}(X)\right| \leq \mathrm{L}(X) \cdot d_r(P, P') \cdot \|\phi\|_r \tag{12}$$

where random variable X is naturally associated with loss, P and P' are two distributions, ρ_ϕ is a Spectral Risk Measure, $\|\phi\|_r = (\int_0^1 \phi^r(\alpha)\mathrm{d}\alpha)^{\frac{1}{r}}$ is the norm[14] of $\phi \in L^r([0,1],\lambda)$ on the standard space $([0,1],\lambda)$ under Lebesgue measure[15] λ and L is a constant.

The left side of the inequality represents the difference between the risk measure of two given distributions. Inequality (12) illustrates that this difference is bounded by two factors: Wasserstein distance $d_r(P, P')$ and the nature of the Spectral Risk Measures themselves, as denoted by $\|\phi\|_r$. To summarize, by calculating Wasserstein metric, there is an upper bound for the change in the risk measure all the time, no matter how the probability measure is being perturbed.

4.1.2 Uncertainty Set Based on Wasserstein Distance and the Size

[14] Norm is a function that assigns a positive length or size to each vector.
[15] Lebesgue measure is used to assign a measurement to subsets of n-dimensional space. For example, when n = 1, 2 or 3, the standard measure is length, area, or volume.

Back to our research question, as discussed in Section 1.1 and Section 2, most current risk metrics are simulated from historical data. To be sure, it is recognized the theoretical limitations of this approach: its projections are directly derived from the distribution of past occurrences and may be irrelevant or even unhelpful if the future is statistically different from the past. And in practice, VaR and other risk models have continually come up short.

In order to simulate the "future" distributions, here we not only consider the individual historical gain and loss, but also include the similar investment's realized distributions. But the new questions arise: how to define similarity and how far we should go?

We can formalize the notion of similarity by the following inequality:

$$d(P, P') \leq c \tag{13}$$

We say two distributions P and P' are similar if they satisfy the relationship (13). Hence, the key parameter here is the radius: c, which describes our sensitivity to the similarity.

It is however difficult to determine c: the radius within similar distributions cannot be set up either too small, making our comparison objectives too few and not comprehensive, or cannot be too large, which will include superfluous investments and cause an over-conservative and meaningless result. In the next section, we purpose a clustering approach to bypass the difficulty of determining c.

4.1.3 Wasserstein-based Clustering

To answer the question about radius r needed in defining the Wasserstein set, we are going to use the clustering method as discussed in Section 3. As shown later, identifying the cluster will automatically imply the radius c. Recall that clustering is the process of construction and subdivision of homogeneous groups of objects in a population in such a way that objects in the same group show a high similarity whereas objects in different groups are typically more dissimilar, where such groups are called "clusters".

We also elaborate different widely-used distance functions to measure the similarities in Section 3.2, including Euclidean distance, Manhattan Distance, etc. However, for our purpose, we need a new distance function: Wasserstein distance (see Section 4.1.1). There are two main reasons to choose it:

(1) Wasserstein distance has a close relationship with risk measure (see in Section 3.2.1). It is known that the results of VaR and CVaR are always sensitive to distribution and parameters decided, and many of them turned out to be disappointing and conservative. But by utilizing Wasserstein distance, as their upper bound metric, more potential extremums can be included and the results can be more robust.

(2) Given that risk measures must be based on the empirical P&L distributions, the function must be able to quantify the distance between distributions, not simply data points or vectors. Recall that Wasserstein distance calculates the geometric discrepancy between distributions by measuring the minimal amount of "work"

needed to move all the mass contained in one distribution onto the other (Justin, 2015). Hence Wasserstein metric, instead of Euclidean or Manhattan distance, is a better way to measure the differences among distributions, including P&L distributions in risk management as well.

We also follow Rachev et al. (2008) and specify the r in Wasserstein distance, where $\mathrm{r} = 2$. Although $\mathrm{r} = 1$ is also useful, $\mathrm{r} = 2$ (also called W_2 distance) is usually stronger (Villani, 2009) and simple to manipulate. Canas and Rosasco (2012) also point out that many unsupervised learning techniques, such as clustering, involve constructing a simple approximation such that W_2 distance is small.

Therefore, the distance function that we use is:

$$d_2(F, \mathrm{G}) = \sqrt{\int_0^1 (F^{-1}(u) - G^{-1}(u))^2 \,\mathrm{du}} \tag{14}$$

After defining distance function, we then need to determine other factors in clustering method. Technically speaking, clustering method can be classified according to data type, algorithm, classification structure, etc. Here, as mentioned, we use Wasserstein metric as the distance measure to quantify similarity between different P&L distributions. Speaking of clustering algorithm, we are going to use the hierarchical[16] clustering (see reason in Section 3.3), which is also the main clustering technique that is heavily used in the literature.

[16] When we say "hierarchical clustering", we refer to agglomerative clustering. The same below.

The following pseudo code is provided for the detailed input, output and algorithm in our framework.

Wasserstein-based Hierarchical Clustering

Input: n stocks' Profit and Loss distributions, each generated from m historical log returns within a certain time window.

Output: Clusters of stocks and dendrogram.

Algorithm:

(1) Turn each input into a singleton, i.e., into a cluster C_i consisting of a single stock, in total there are n clusters.

(2) For each pair of clusters C_1, C_2, calculate their Wasserstein distance $d_2(C_1, C_2)$, as shown in equation (14).

(3) Find and merge the most similar pair (C_i, C_j), which take the smallest Wasserstein distance, into a single cluster to form the next clustering.

(4) Continue step (2) and (3) until the there is only one cluster, which contain all the input stocks.

To summarize, regardless of which algorithm we choose, one of the main reasons for choosing the clustering approach is that it can help us to bypass the problem of the radius between similar distributions. Rather than determining a specific figure r, the question turns to defining how many clusters we should have. As the important parameter changes

from radius to cluster number, clustering will automatically mate the selection of radius within different clusters. Compared to a relatively abstract radius, decision of cluster number is much more intuitive: according to sector in stock market, industry in company classification, etc.

4.1.4 Method Summary

In our application, the method follows these steps:

Step 1. Generate P&L distributions from investments.

Step 2. Calculate the Wasserstein distance between each two distributions.

Step 3. Determine cluster number.

Step 4. Use the Wasserstein-based clustering and cluster number to get similar distribution sets (see Section 4.1.3 for the pseudo code).

Step 5. Within each set, calculate the risk measures for each distribution and use the highest result here to denote the possible highest risk for any distribution in the set.

Step 6. Verify the validation by backtesting, see Section 4.2.

4.2 Validation Approach: Backtesting

After generating results, the next task is to verify the accuracy and reliability of different risk measures generated from our method. Backtesting is the comparison of the actual gain and loss with the estimated results of a market risk measurement method, which

is a set of statistical procedures designed to check if the real losses, observed ex post, are in line with previous forecasts (Jorion, 2007). Backtesting allows us to address the question of whether a given estimation procedure produces credible risk-measure estimates. If the estimation result is similar to the actual result, it indicates that the accuracy and reliability of the risk measurement method or model are high; if the difference between the two is large, it indicates that the accuracy or reliability of the risk measurement method or model is low, or there is a problem with the hypothesis of the backtesting test. Test result between the two cases implies that there is a problem with the risk measurement method or model, but the conclusion is uncertain.

4.2.1 Value at Risk Backtesting

For VaR, since 1990s, several tests have been proposed which can be used to measure the accuracy of a proposed VaR model. Although the details of those tests are different, many of these tests focus on the straightforward comparison between the reported VaR and realized gain or loss by counting the number of exceedances, which is the number of realized losses that exceeded the predicted VaR level.

To be more specific, consider the event that the loss on a portfolio exceeds its estimated VaR_α at certain confidence level α. Denoting the profit or loss on the portfolio over a fixed time interval as x_t then define the "hit" function, or the fail rate, as follows:

$$I_{t+1}(\alpha) = \begin{cases} 1, if\ x_{t+1} \leq VaR_\alpha \\ 0, if\ x_{t+1} > VaR_\alpha \end{cases}$$

(15)

so that the hit function sequence, i.e., $(1, 1, 0, 1, 1, \dots, 0)$, tallies whether or not a loss in excess of the reported VaR_α has been realized.

Christoffersen (1998) points out the problem of determining the accuracy of VaR can be reduced to the problem of testing the hit sequence as described. The most common method here is to test the unconditional coverage property, assuming independence of extreme events is tested individually:

According to the definition, each potential exceedance is a Bernoulli distributed random variable. And the probability of realizing a loss in excess of the reported VaR_α should be $\alpha \times 100\%$ or in terms of the notation, $Pr = (I_{t+1}(\alpha) = 1) = \alpha$. If it is the case that losses in excess of the reported VaR_α occur more frequently than $\alpha \times 100\%$ of the time then this would suggest that the reported VaR_α measure systematically understates the portfolio's actual level of risk, while too few violations would oppositely signal an over-conservative VaR measure. As a result, especially for some early propositions, they are concerned with whether or not the reported VaR_α is violated more (or less) than $\alpha \times 100\%$ over a given span of time.

4.2.2 Conditional Value at Risk Backtesting

One of the first and still most frequently used tests for the CVaR is the exceedance residual backtesting of (McNeil and Frey, 2000). Their approach is distribution-free and based on the CVaR residuals that exceed the VaR. Inspired by (15), firstly we can get:

$$\mathrm{E}\left[\frac{l_{t+1}}{CVaR_{\alpha,t}} - 1 \middle| l_{t+1} > VaR_{\alpha,t}\right] = 0 \tag{16}$$

where l_t is the loss of the portfolio.

After testing VaR, define VaR violation $\{l_{t+1} > VaR_{\alpha,t}\}$, the indicator function of an α–exception. The backtest statistic for CVaR here is defined as a violation residual:

$$K_{t+1} = (\frac{l_{t+1} - CVaR_{\alpha,t}}{CVaR_{\alpha,t}})I_{t\{l_{t+1}>VaR_{\alpha,t}\}} \tag{17}$$

The backtesting procedure analyses the empirical difference between the next period's loss l_{t+1} and $CVaR_{\alpha,t}$, which is the expected shortfall at time t, to the forecast $CVaR_{\alpha,t}$, conditional on the fact that l_{t+1} exceeds the VaR at time t, $VaR_{\alpha,t}$. Thus, the backtest examines the distance / relation between the CVaR-forecasts and realized payoffs. If there is a VaR violation, then $I_{t\{l_{t+1}>VaR_{\alpha,t}\}} = 1$, K_{t+1} compares the actual violation size with the expected size given by $CVaR_{\alpha,t}$; if there is no VaR violation, then $I_{t\{l_{t+1}>VaR_{\alpha,t}\}}$ and K_{t+1} will both be 0.

Under the null hypothesis (H_0), McNeil and Frey (2000) argue that the sequence of violation residual-K_t forms an independent identically distributed (i.i.d) process with mean 0 and variance 1. If the mean is negative, it means the CVaR is overestimated.

4.2.3 Spectral Risk Measures Backtesting

The accepted methods to backtest Spectral Risk Measures are actually elusive. Here we follow the Coverage Rate method, with a Z-test for a discretized version of SRMs, proposed by Costanzino and Curran (2015). The reason why we choose this backtesting method is that it relies on an appropriate extension of the VaR breach indicator to the case of SRMs. The resulting new breach indicator (18) takes into account the severity of the breach (i.e., losses beyond the VaR level) and is a continuous variable rather than discrete.[17]

The backtest starts with the failure rate X_ϕ^N in analogy with the VaR's. For an admissible spectrum ϕ, let $X_\phi^{(i)} \in [0,1]$ be defined by

$$X_\phi^{(i)}(\phi) = \int_0^1 \phi(\mathrm{p})\mathbf{1}_{\{l_i \le VAR_i(p)\}} dp \tag{18}$$

Then define the Spectral Risk Measures failure rate X_ϕ^N for admissible risk spectrum ϕ as

$$X_\phi^N(\phi) = \frac{1}{N}\sum_{i=1}^{N} X_\phi^{(i)}(\phi) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 \phi(\mathrm{p})\mathbf{1}_{\{l_i \le VAR_i(p)\}} dp = \frac{1}{N}\sum_{i=1}^{N}\int_{VAR_i^{-1}(l_i)}^1 \phi(\mathrm{p})dp$$
$$= 1 - \frac{1}{N}\sum_{i=1}^{N}\Phi(VAR_i^{-1}(l)) \tag{19}$$

[17] Costanzino, N. & Curran, M., 2018. A Simple Traffic Light Approach to Backtesting Expected Shortfall. Risks, 6(1), pp.3.

where $\Phi' = \phi$.

Then the null-hypothesis for the Spectral Risk Measures Coverage Test is that $\{X_\phi^{(i)}\}_{i=1}^N$ are i.i.d. with bounded mean μ_ϕ and variance σ_ϕ^2 given by (20) and (21) respectively.

$$\mu_\phi = \mathbb{E}\left[X_\phi^N(\phi)\right] = \int_0^1 \phi(\mathrm{p}) p dp \tag{20}$$

$$\sigma_\phi^2 = \mathbb{V}\left[X_\phi^N(\phi)\right] = \frac{1}{N}\left(2\int_0^1 \int_0^p \phi(\mathrm{p})\phi(\mathrm{q}) dp dq - \left(\int_0^1 \phi(\mathrm{p}) p dp\right)^2\right) \tag{21}$$

And the expected value of the average over N trading days is equal to the expected value of a single trading day. For large enough N, X_ϕ^N is approximately normal and therefore admits a Z-test. The Z-score Z_ϕ^N can be defined by

$$Z_\phi^N = \frac{\hat{X}_\phi^N(\phi) - \mu_\phi}{\sigma_\phi} \tag{22}$$

We can then conduct a Z-test for the Spectral Risk Measures failure rate to verify whether it is normally distributed under different confidence levels with large samples.

5. Results

5.1 Data and Overview

Our data source is Yahoo Finance, where we take 10-year daily stock price of 120 stocks (2519 observations individually) from Standard & Poor's 500 Index (S&P 500) and calculate their daily log return. From that, we construct the Profit and Loss (P&L) distributions for 120 stocks, and each 20 stocks are from the same sector according to Global Industry Classification Standard (GICS). Therefore, regarding to sector, there are 6 similar groups in total: Consumer Discretionary, Energy, Health Care, Industrials, Information Technology and Real Estate. Time window here is a 10-year period, from Apr 20, 2008 to Apr 20, 2018.

We first test the sector-based categorization by randomly choosing 4 stocks from each sector and make a comparison among stocks in the same sector (see Fig. 9). It can be observed that although the Profit & Loss distributions in the same column (sector) share something similar, but their left tails diverge considerably.

For example, for the first column-Health Care, the middle of the distributions is somehow similar, while the loss part of the third stock is much more extreme, and its right tail, denoting the profit, is also more obvious than other stocks in the same sector. It means that the price of this stock fluctuates wildly, bringing considerable returns as well as losses. The same situation also occurs in the last stock in Industrials sector and second stock in

Real Estate secotr. They have greater price volatility and associated risks than other more concentrated distributions.

When it comes to the extreme losses, we refer to the last column-Consumer Discretionary, where the worst-case risks vary from -0.1 to over -0.2. It is conceivable that if people use these four stocks as a reference to estimate their risks, the result will be very different from the reality.

The first result suggests that sector categorization is not risk-orientated, suggesting that stocks in the same sector do not share the same risk profiles.

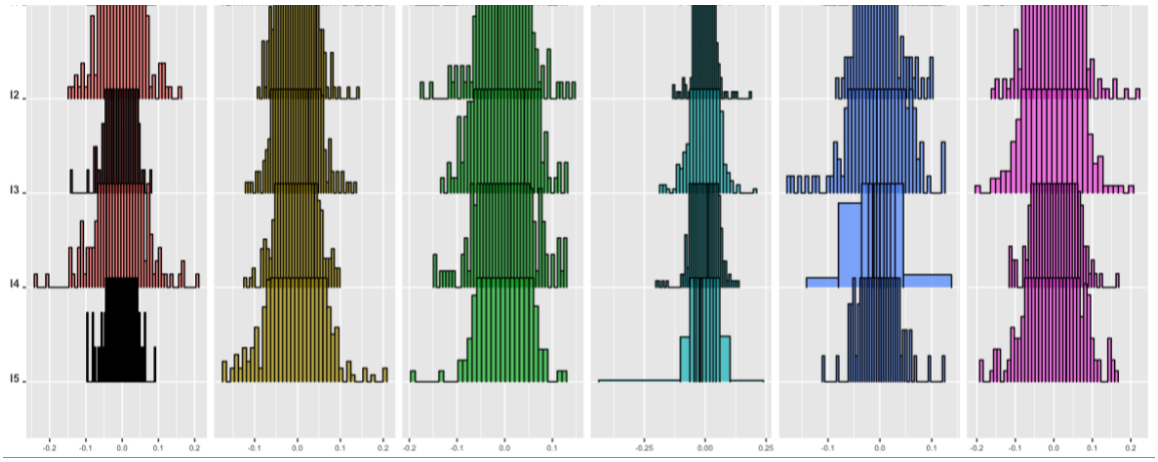


Fig. 9 P&L distributions among sectors

Sector from left to right: Health Care, Industrials, Information Technology, Energy, Real Estate and Consumer Discretionary

Next, we construct the Wasserstein-based clustering among those 120 distributions. To compare different approaches, we apply two kinds of hierarchical clustering (see reason in Section 3.3.2), based on Euclidean distance and Wasserstein distance for 120 stocks' log

return distributions. We also test the relationship between Wasserstein distance and risk measures (CVaR and SRMs here).

After getting clustering results, we first calculate 90%, 95% and 99% VaR and CVaR of each stock, and present the results by the following categorizations:

(1) Sectors

(2) Clusters based on Euclidean distance

(3) Clusters based on Wasserstein distance

Last but not least, we use backtesting, discussed in Section 4.2 to verify our new Wasserstein-based clustering method and compare the results to the original historical estimation.

5.2 Numerical Results

5.2.1 Wasserstein-based Clustering Categorization

We first present the result of hierarchical clustering based on Wasserstein distance in Table. 2. In order to compare the result with their original sectors, here we determine cluster number is 6. The stock is presented is the form of the ticker symbol, and the number attached to each symbol informs the stock's sector, where number 1-6 represents the stock belongs to Consumer Discretionary, Energy, Health Care, Industrials, Information Technology and Real Estate, respectively.

The clustering result shows that Cluster 1 and 2 both include over 30 stocks from different 6 sectors, meaning that despite the different sectors, the results in clustering still show they have close Wasserstein distances. For cluster 3, most members are from Consumer Discretionary, but it also includes several stocks from other 4 sectors, which indicates those stocks from different sectors may share some implicit similarities. Cluster 4, 5 and 6 are quite special, and we can see that the stocks from sector Information Technology and Real Estate show very different profile when classified by Wasserstein distance.

Cluster	**Stock Symbol**
Cluster 1 (35 stocks)	AAP1, AMZN1, BWA1, KMX1, CHTR1, CMG1, DISCA1, DISH1, EXPE1
	APA2, EOG2, EQT2, MPC2
	ALXN3, BIIB3, BSX3, CNC3
	JEC4, KSU4, PWR4
	EA5, FFIV5, FB5, HPQ5, JNPR5
	EQIX6, EXR6, FRT6, HCP6, MAA6, PSA6, O6, REG6, SBAC6, SPG6
Cluster 2 (57 stocks)	APTV1, AZO1, CCL1, CMCSA1, DRI1, DISCK1, DG1, DLTR1
	CVX2, COP2, XOM2, KMI2
	ABT3, ABBV3, AET3, A3, AGN3, ABC3, AMGN3, ANTM3, BAX3, BDX3, BMY3, CAH3, CELG3, CERN3, CI4

	HII4, INFO4, ITW4, IR4, JBHT4, JCI4, LLL4, LMT4, NLSN4, NSC4, NOC4, PCAR4, PH4, PNR4, RTN4, RSG4
	EBAY5, FIS5, FISV5, FLIR5, IT5, GPN5, HRS5, HPE5, INTC5, IBM5, INTU5, KLAC5
	ESS6, IRM6
Cluster 3 (18 stocks)	BBY1, CBS1, DHI1
	APC2, ANDV2, BHGE2, COG2, XEC2, CXO2, DVN2, HAL2, HP2, HES2, MRO2, NOV2
	ALGN3
	MAS4
	IPGP5
Cluster 4 (7 stocks)	DRE6, HST6, KIM6, MAC6, PLD6, SLG6, VTR6
Cluster 5 (2 stocks)	DXC5, LRCX6
Cluster 6 (1 stock)	GGP5

Table. 2 Wasserstein-based clustering result

5.2.2 Spectral Risk Measure Comparisons among Three Categorizations

In this section, we calculate risk measures, as stated in Section 2. To be comprehensive and consistent with inequality (12), where ρ_ϕ is a Spectral Risk Measure, we present the Spectral Risk Measures comparisons based on a discrete spectrum:

$$\text{SRMs} = \phi_1 * 90\%\, CVaR + \phi_2 * 95\%\, CVaR + \phi_3 * 99\%\, CVaR \tag{23}$$

We choose three spectrums, each one assigning different weights to 90%, 95% and 99% CVaR. To be more specific,

$$\phi_I = \begin{cases} 0.3, & at\ 90\%\ CVaR,\ \ i = 1 \\ 0.3, & at\ 95\%\ CVaR,\ \ i = 2 \\ 0.4, & at\ 99\%\ CVaR,\ \ i = 3 \end{cases}$$

$$\phi_{II} = \begin{cases} 0.1, & at\ 90\%\ CVaR,\ \ i = 1 \\ 0.3, & at\ 95\%\ CVaR,\ \ i = 2 \\ 0.6, & at\ 99\%\ CVaR,\ \ i = 3 \end{cases}$$

and

$$\phi_{III} = \begin{cases} 0.1, & at\ 90\%\ CVaR,\ \ i = 1 \\ 0.2, & at\ 95\%\ CVaR,\ \ i = 2 \\ 0.7, & at\ 99\%\ CVaR,\ \ i = 3 \end{cases}$$

To compare the results among three categorizations, we provide three boxplots (Fig. 10 - 12) by grouping the abovementioned SRM according to the categorizations stated in the overview section (sectors; clusters based on Euclidean distance and clusters based on Wasserstein distance). Each box demonstrates the risk measure calculation within that category and the height of each box informs the degree of SRMs separation.

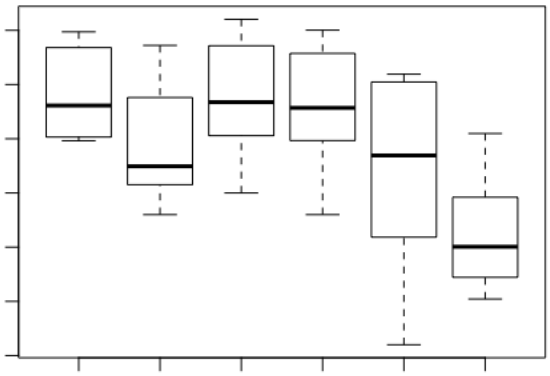


Fig. 10 (a) SRMs comparison among 6 sectors

$$\phi_I = \begin{cases} 0.3, & at\ 90\%\ CVaR \\ 0.3, & at\ 95\%\ CVaR \\ 0.4, & at\ 99\%\ CVaR \end{cases}$$

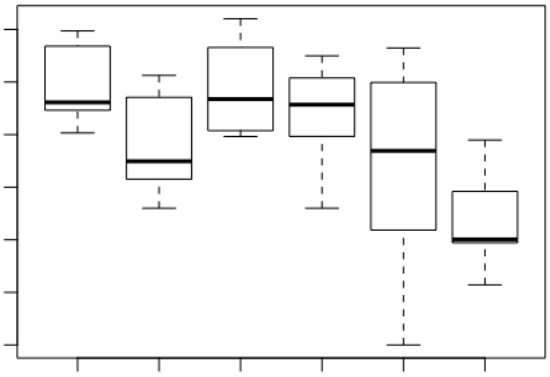


Fig. 10 (b) SRMs comparison among 6 sectors

$$\phi_{II} = \begin{cases} 0.1, & at\ 90\%\ CVaR \\ 0.3, & at\ 95\%\ CVaR \\ 0.6, & at\ 99\%\ CVaR \end{cases}$$

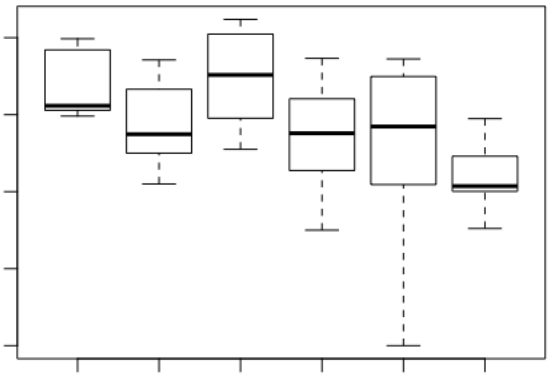


Fig. 10 (c) SRMs comparison among 6 sectors

$$\phi_{III} = \begin{cases} 0.1, & at\ 90\%\ CVaR \\ 0.2, & at\ 95\%\ CVaR \\ 0.7, & at\ 99\%\ CVaR \end{cases}$$

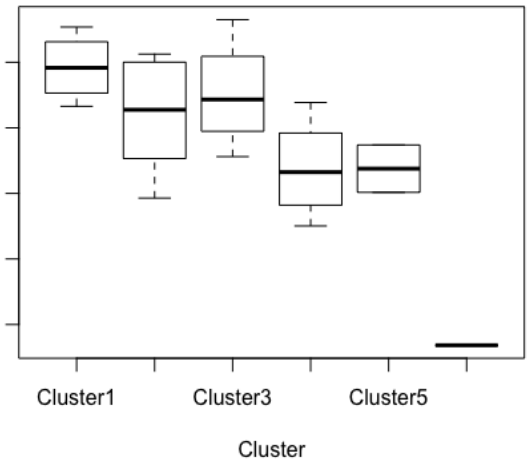


Fig. 11 (a) SRMs comparison among 6 clusters based on Euclidean distance clustering

$$\phi_{I} = \begin{cases} 0.3, & at\ 90\%\ CVaR \\ 0.3, & at\ 95\%\ CVaR \\ 0.4, & at\ 99\%\ CVaR \end{cases}$$

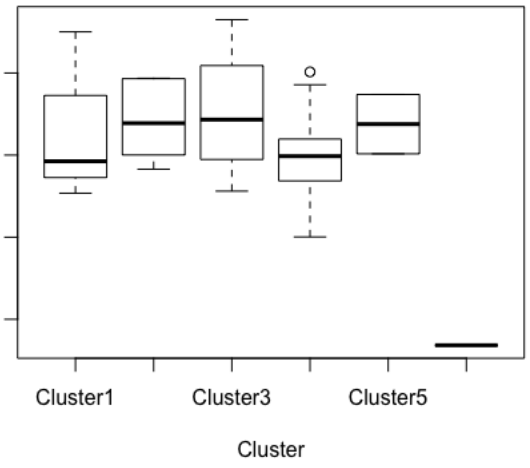


Fig. 11 (b) SRMs comparison among 6 clusters based on Euclidean distance clustering

$$\phi_{II} = \begin{cases} 0.1, & at\ 90\%\ CVaR \\ 0.3, & at\ 95\%\ CVaR \\ 0.6, & at\ 99\%\ CVaR \end{cases}$$

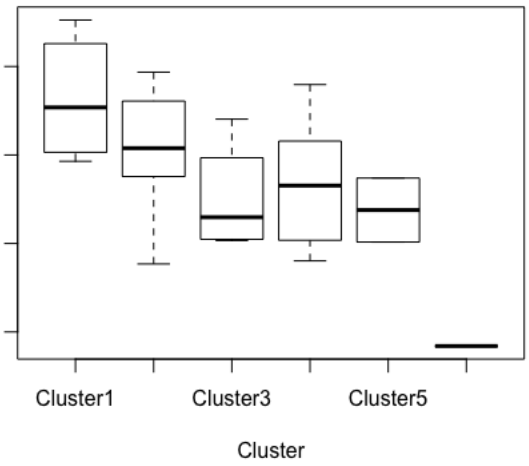


Fig. 11 (c) SRMs comparison among 6 clusters based on Euclidean distance clustering

$$\phi_{III} = \begin{cases} 0.1, & at\ 90\%\ CVaR \\ 0.2, & at\ 95\%\ CVaR \\ 0.7, & at\ 99\%\ CVaR \end{cases}$$

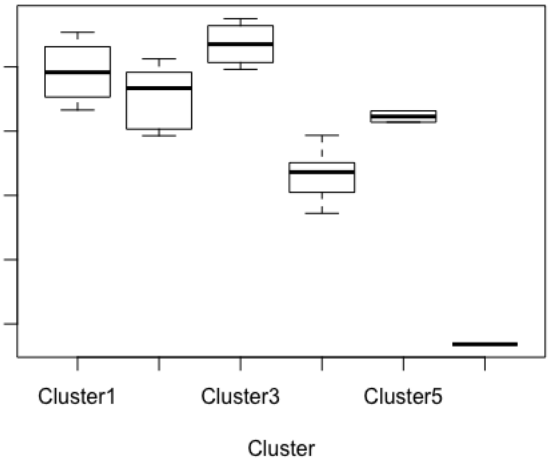


Fig. 12 (a) SRMs comparison among 6 clusters based on Wasserstein distance clustering

$$\phi_{I} = \begin{cases} 0.3, & at\ 90\%\ CVaR \\ 0.3, & at\ 95\%\ CVaR \\ 0.4, & at\ 99\%\ CVaR \end{cases}$$

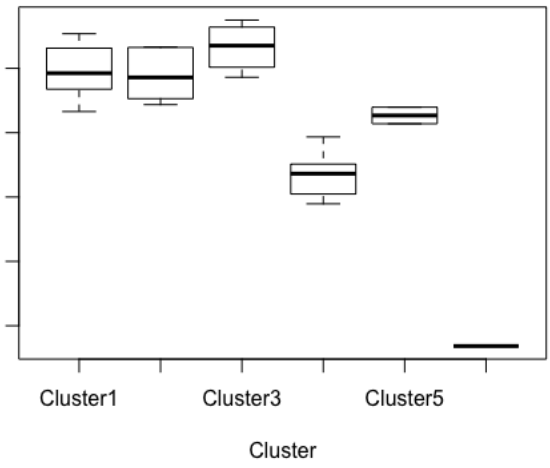


Fig. 12 (b) SRMs comparison among 6 clusters based on Wasserstein distance clustering

$$\phi_{II} = \begin{cases} 0.1, & at\ 90\%\ CVaR \\ 0.3, & at\ 95\%\ CVaR \\ 0.6, & at\ 99\%\ CVaR \end{cases}$$

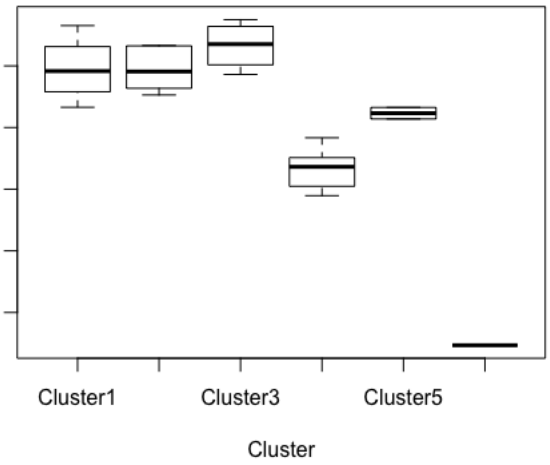


Fig. 12 (c) SRMs comparison among 6 clusters based on Wasserstein distance clustering

$$\phi_{III} = \begin{cases} 0.1, & at\ 90\%\ CVaR \\ 0.2, & at\ 95\%\ CVaR \\ 0.7, & at\ 99\%\ CVaR \end{cases}$$

By looking at Fig. 10, Fig. 11 and Fig. 12, we see that the approximate average height of the boxes in Fig. 12 (a), (b) and (c) is significantly smaller than that in Fig. 10 and 11.

It is known that box plot demonstrates the dispersion of a set of data, so the smaller the length of the box, the more concentrated the data, the no outlier.

That is to say, within the same cluster classified by Wasserstein distance (as shown in Fig. 12 (a), (b) and (c)), the values of SRMs in the same cluster are more concentrated and the dispersion is smaller, which indicates that within distribution sets identified by Wasserstein distance, risk measure results are more similar, comparing to the sector groups and distribution sets based on Euclidean distance.

To summarize, given that from sector-classification to Wasserstein-based clustering, the risks between "similar" stocks are indeed more similar, distributions with smaller Wasserstein distances have more similar risk measures. It also means that compared to classical sector classification and Euclidean distance clustering, Wasserstein-based clustering yields a more valuable result in terms of Spectral Risk Measures.

5.2.3 Verification of the Relationship between Wasserstein Distance and Spectral Risk Measures

To verify the relationship (see inequality (12) in Section 4.1.1) between Wasserstein metric and Spectral Risk Measures, we calculate the Wasserstein distance between each two stocks and the difference of 95% Conditional Value at Risk between each two stocks. If (12) holds, we should be able to observe that the difference in CVaR between any two

stocks is bounded by their Wasserstein distance. The line chart (Fig. 14) of Euclidean distance and difference of CVaR is also listed for comparison purpose.

From Fig. 13, it can be seen that there is an implicit linear relationship between CVaR difference and Wasserstein distance: as the Wasserstein distance increases, the value of CVaR difference and its upper limit also increase (as illustrated by the dashed line). But when it comes to Fig. 14, the relationship between Euclidean distance and CVaR is less clear, and we cannot draw a line between the Euclidean distance and risk measure difference. Therefore, by comparing Fig. 13 to Fig. 14, we verify the relationship between Wasserstein distance and SRMs, as shown in inequality (12).

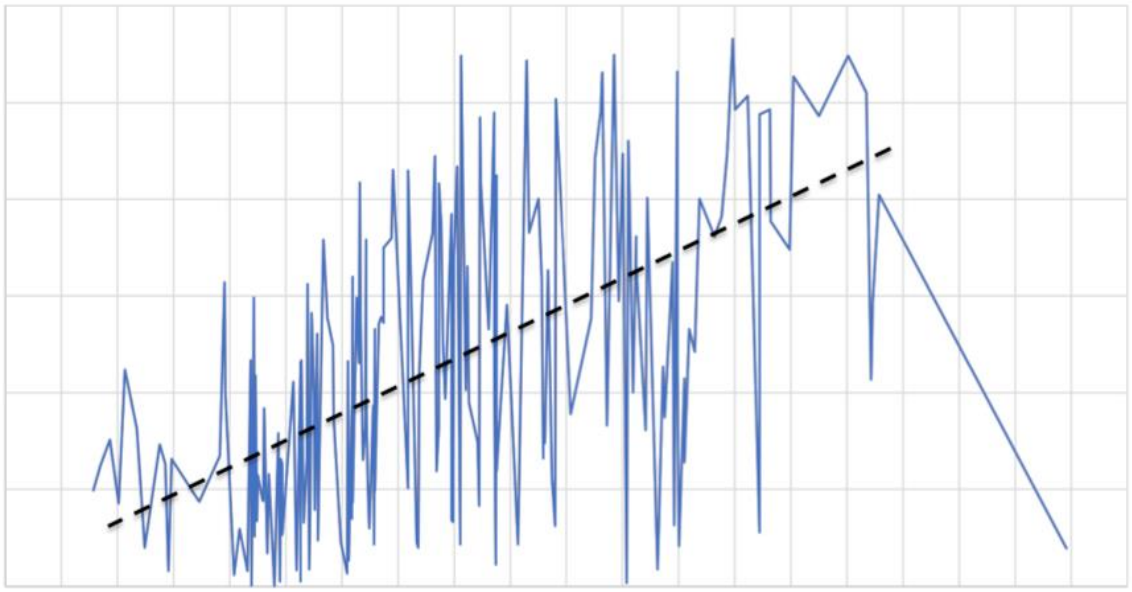


Fig. 13 95% CVaR difference and Wasserstein distance

Vertical coordinate: 95% CVaR; Horizontal coordinate: Wasserstein distance

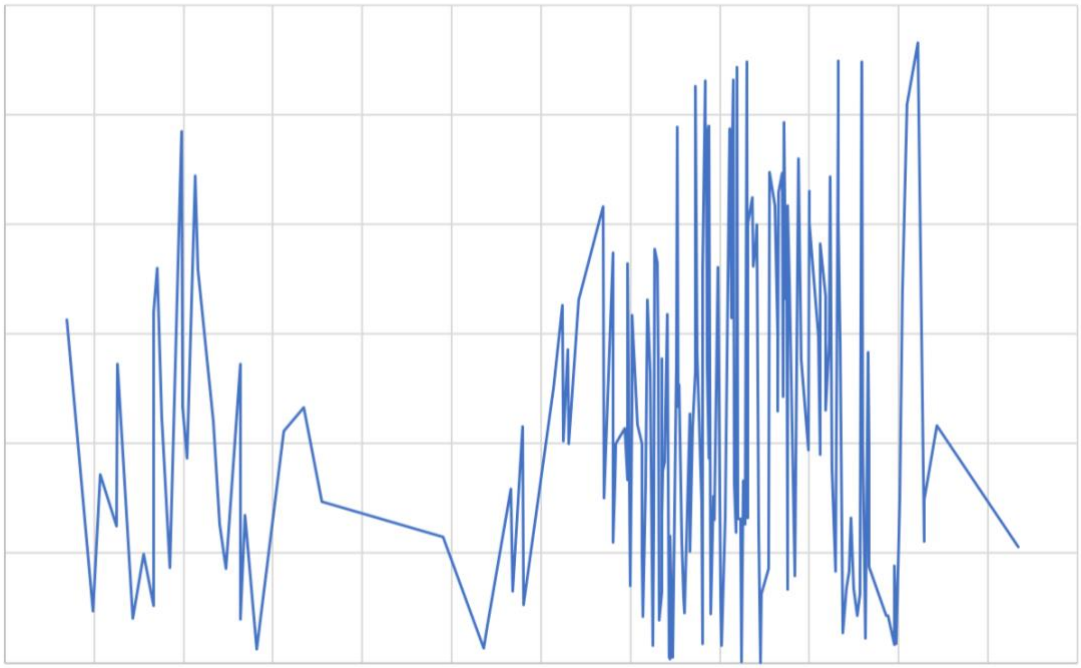


Fig. 14 95% CVaR difference and Euclidean distance

Vertical coordinate: 95% CVaR; Horizontal coordinate: Euclidean distance

5.2.4 Backtesting of Value at Risk, Conditional Value at Risk and Spectral Risk Measures

In this section, we use the backtesting, as stated in Section 3.5, to verify the validation of our new approach to measure risk. Recall that for VaR violation rate and CVaR backtesting result introduced in Section 3.5, the closer their values are to 0, the more accurate and effective the prediction of VaR or CVaR is.

In practice, within each Wasserstein-based cluster, for each day, we first get the largest value of 90%, 95%, 99% VaR to denote the new worst possible Value at Risk in each cluster. Then we use the hit function backtesting method to calculate the failure rate based

on new “worst-case” VaR for all the stocks in the same cluster. For all the 120 stocks, we then get 120 backtesting results for the new VaR from our method, and we present the average of those 120 results, called “Average of Worst VaR within Cluster” in Table. 3. For comparison purpose, we also do backtesting for the original VaR at 90%, 95%, 99%, which is conducted from each investment’s P&L distribution only. We present the average result of that backtesting by “Average of Each Stock's Own VaR” in Table. 3.

VaR Confidence Level	Average of Each Stock's Own VaR	Average of Worst VaR within Cluster
90%	0.157366070	0.052631578
95%	0.106629980	0.02671875
99%	0.018991494	0.005400000

Table. 3 VaR Backtesting Comparison

As mentioned in Section 4.2.1, for VaR backtesting result, the closer to $1-\alpha$, the better, meaning the failure rate is closer to the theoretical level $(1-\alpha)$. And the most ideal result is $1-\alpha$, meaning the probability of realized payoff that is below the estimated VaR is exactly the confidence level α. In Table. 3, we can see that the “Average of Each Stock's Own VaR” is always above $1-\alpha$ (10%, 5% and 1% here), so the initial VaR results underestimate the actual losses. But in terms of “Average of Worst VaR within Cluster”, there are huge decrease from the original results; and at 95% and 99% confidence levels, our method conducts safer results according to the theoretical value of $1-\alpha$.

Next, recall that in inequality (12), the relationship is between Wasserstein distance and Spectral Risk Measures, therefore it is necessary to verify the validation for SRMs results. We first choose the CVaR, which is the simplest form (see Section 2.4) of SRMs, at the confidence level of 90%, 95%, 99%. The process is similar to VaR:

After calculating the CVaR for each stock, we use the largest absolute amount of 90%, 95%, 99% CVaR to denote the new worst-case Conditional Value at Risk for all stocks within the same cluster. Then we use the violation residual backtesting method to calculate the exception ratio (see equation (17)) based on new worst-case CVaR and realized gain and loss for each stock. For all the 120 stocks, we then get 120 backtesting results for the new CVaR from our method, and we present the average of those 120 results, called "Average of Worst CVaR within Cluster" in Table. 4. Also For comparing, we do backtesting for the original CVaR at 90%, 95%, 99% confidence level, which is again from each investment's individual P&L distribution only. We present the average result of that backtesting by "Average of Each Stock's Own CVaR" in Table. 4.

CVaR Confidence Level	Average of Each Stock's Own CVaR	Average of Worst CVaR within Cluster
90%	0.12010339	-0.08589021
95%	0.004195534	0.003243282
99%	0.000628441	-0.000388488

Table. 4 CVaR Backtesting Comparison

For CVaR backtesting, there is no intuitive failure rate, which can be used to compare the realized violation directly. Recall Section 4.2.2, we use the residual ratio to examine the distance between the CVaR-forecast and realized payoff. Ideally, such ratios should be independent identically distributed (i.i.d), with mean 0 and variance 1. In practice, if the risk measure estimation is unbiased, the CVaR backtesting result should be 0. In other words, the closer the backtesting CVaR is to 0, the better, indicating the realized loss closer to the average (CVaR).

As shown in Table. 4, for each confidence level, the backtesting result from "Average of Worst CVaR within Cluster" is significantly closer to 0, compared to "Average of Each Stock's Own CVaR". That is to say, in comparison to the CVaR of each stock itself, our approach can more accurately predict the future risks, especially in extreme situations (at 10%, 5% and 1% probabilities). The only problem with our method is that at 90% and 99% confidence level, the CVaRs conducted from Wasserstein-based clustering tend to be over-conservative (the backtesting results are negative). There is always a tradeoff between conservatism and asset efficiency as discussed in Section 4.1.2. Here we sacrifice some flexibility, but our results are far more robust and safer than the original CVaR.

To be more comprehensive, lastly, we present the backtesting result for other forms of Spectral Risk Measures. The first type of SRMs that we use is as followed, where $\phi(\gamma)$ is a discrete function (the same as Section 5.2.2)

$$\text{SRMs} = \phi_1 * 90\%\, CVaR + \phi_2 * 95\%\, CVaR + \phi_3 * 99\%\, CVaR \tag{24}$$

There are again three amounts of ϕ_i, where

$$\phi_I = \begin{cases} 0.3, & at\ 90\%\ CVaR,\ \ i = 1 \\ 0.3, & at\ 95\%\ CVaR,\ \ i = 2 \\ 0.4, & at\ 99\%\ CVaR,\ \ i = 3 \end{cases}$$

$$\phi_{II} = \begin{cases} 0.1, & at\ 90\%\ CVaR,\ \ i = 1 \\ 0.3, & at\ 95\%\ CVaR,\ \ i = 2 \\ 0.6, & at\ 99\%\ CVaR,\ \ i = 3 \end{cases}$$

and

$$\phi_{III} = \begin{cases} 0.1, & at\ 90\%\ CVaR,\ \ i = 1 \\ 0.2, & at\ 95\%\ CVaR,\ \ i = 2 \\ 0.7, & at\ 99\%\ CVaR,\ \ i = 3 \end{cases}$$

Recall the definition of SRMs, from ϕ_I to ϕ_{III}, we assign higher weights to 99% CVaR, suggesting a more risk averse towards larger losses. The data processing is similar to CVaR backtesting, while here three weighted average CVaRs are needed. After calculate the 90%, 95%, 99% CVaR, we first refer to the equation (24) and corresponding $\phi(\gamma)$ to obtain $SRMs_I$, $SRMs_{II}$ and $SRMs_{III}$. And the results get here are used for backtesting traditional SRMs, which are from the single distribution only. In Table. 5, we show the average of those results, "Original Average Z_ϕ^N", of 120 stocks.

To do the backtesting for our method, in one cluster, we use the largest absolute amount of $SRMs_I$, $SRMs_{II}$ and $SRMs_{III}$, representing the new worst-case SRMs for the

stocks in the same cluster. Then we calculate the μ_ϕ and Z score in equation (22). We then redo the process for all those 6 clusters and calculate the average of those backtesting results, named "Average Z_ϕ^N from Clustering" in Table. 5.

ϕ_i for 90%, 95%, 99% CVaR	Original Average Z_ϕ^N from Clustering	Average Z_ϕ^N from Clustering
0.3, 0.3, 0.4	1.607458604	0.942374685
0.1, 0.3, 0.6	1.391368942	0.881383410
0.1, 0.2, 0.7	1.309596727	0.602471204

Table. 5 Spectral Risk Measures Backtesting Comparison

Last step is to compare the Z scores to one-side Z table. The "Original Average Z_ϕ^N from Clustering" shows that we can reject the null hypothesis (the $\{X_\phi^{(i)}\}_{i=1}^N$ sequence is i.i.d) at 95%, 90% and 90% confidence level when $\phi = \phi_I, \phi_{II}, \phi_{III}$, respectively. While for "Average Z_ϕ^N from Clustering", we cannot reject the null hypothesis (H_0) at 95% confidence level.

According to Costanzino and Curran (2015), the Spectral Risk Measure failure rate is a continuous variable with value between 0 and 1 and proved to be independent and identically distributed (i.i.d). On top of that, when using the above Z test, the more we can't reject the null hypothesis, the closer Spectral Risk Measure failure rate $\{X_\phi^{(i)}\}_{i=1}^N$ is to i.i.d, and the more accurate SRMs are. In Table. 5, the Z scores show that we cannot reject H_0, suggesting that we should accept H_0, $\{X_\phi^{(i)}\}_{i=1}^N$ sequence is i.i.d in our Wasserstein-based

clustering method, which is consistent with the theoretical SRMs failure rate values. In terms of the "Average Z_{ϕ}^{N} from Clustering", the H_0 should be rejected, indicating a higher-level severity of the SRMs failure in the original method.

6. Conclusion and Direction of Future Research

In our thesis, our main contribution is that we propose a more robust approach to measure market risk given the issue of distribution ambiguity. Our method is motivated by the distribution sets, where investments' Profit and Loss distributions in the same set should share the same, or at least similar, risk profiles. After verifying the shortcoming of traditional categorization-sector, we take advantage of two desirable properties of Wasserstein metric and introduce a Wasserstein-based clustering method for risk measurement. The numerical results generally align with our goal which is to provide a better and more robust prediction of potential extreme risks. We also use the backtesting to exam validation and get consistent results.

One major limitation of the thesis comes from the data source. Here we use Yahoo Finance, which has a robust data and content capabilities and displays and the vast majority of information appears to be accurate and timely. But finance data from Yahoo Finance still seem to have a few inaccuracies, i.e., stock prices with dividends, splits and acquisitions. As a result, data in Section 5.1 could be unclear and even slightly erroneous. One way to address this limitation is using data from Bloomberg or other professional financial datasets to ensure the accuracy and reliability.

Another potential limitation is the choice of data. We only use stocks from S&P 500, but in reality, there are considerable listed companies and private companies. In order to be comprehensive, we may need to extend the choice of data and include different types of

companies.

Speaking of direction of future research, firstly, risk measures of individual stock can be extended to portfolio. The motivation is that financial institutions manage portfolios, so risk measurement of portfolios should be more appealing and valuable. Secondly, we only compare our Wasserstein-based clustering results with the sector-based categorization, regardless of industry or sub-industry. If a more detailed categorization is used, the similarity of stocks in the same category may be higher and the final results may change, which is therefore worth studying. Last but not least, after we have come up with the new risk-oriented classifications, we can study the factors behind. Excluding industry factors, there could be other macro or micro factors that affect market risk metrics of stocks, which should be emphasized as well.

Reference

[1] Acerbi, C. & Tasche, D., 2002. On the coherence of expected shortfall. Journal of Banking and Finance, 26(7), pp.1487–1503.

[2] Acerbi, C. & Székely, B., 2015. Back-testing expected shortfall. Insurance Risk, pp.35–40.

[3] Acerbi, C., 2002. Spectral measures of risk: A coherent representation of subjective risk aversion. Journal of Banking and Finance, 26(7), pp.1505–1518.

[4] Ajiboye, A.R. & Olufadi, H.I., 2018. Performance evaluation of distance metrics in the clustering algorithms. International Journal of Software Engineering and Computer Systems, 4(2), pp.38–48.

[5] Artzner, P. et al., 1999. Coherent measures of risk. Mathematical Finance, 9(3), pp.203–228.

[6] Balbás, A., Garrido, J. & Mayoral, S., 2009. Properties of distortion risk measures. Methodology and Computing in Applied Probability, 11(3), pp.385–399.

[7] Bigot, J. et al., 2018. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. Electronic Journal of Statistics, 12(2), pp.2253–2289.

[8] Bilgili, M. & Mulvey, John M., 2009. Clustering techniques and multi -regime stochastic optimization with applications in finance, pp.ProQuest Dissertations and Theses.

[9] Bini, B.S. & Mathew, T., 2016. Clustering and regression techniques for stock prediction. Procedia Technology, 24, pp.1248–1255.

[10] Brandtner, M., 2013. Conditional Value-at-Risk, spectral risk measures and (non-)diversification in portfolio selection problems – A comparison with mean–variance analysis. Journal of Banking and Finance, 37(12), pp.5526–5537.

[11] Campbell, S., 2006. A review of backtesting and backtesting procedures. The Journal of Risk, 9(2), pp.1–17.

[12] Cerreia-Vioglio et al., 2013. Ambiguity and robust statistics. Journal of Economic Theory, 148(3), pp.974–1049.

[13] Chen, J., 2014. Measuring market risk under the Basel accords: VaR, stressed VaR, and expected shortfall. Aestimatio, (8), pp.184–201.

[14] Cont, R., Deguest, R. & Scandolo, G., 2010. Robustness and sensitivity analysis of risk measurement procedures. Quantitative Finance, 10(6), pp.593–606.

[15] Costanzino, N. & Curran, M., 2015. Backtesting general spectral risk measures with application to expected shortfall. The Journal of Risk Model Validation, 9(1), pp.21–31.

[16] Costanzino, N. & Curran, M., 2018. A simple traffic light approach to backtesting Expected Shortfall. Risks, 6(1), p.2–9.

[17] Delage, E. & Yinyu, Y., 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research, 58(3), pp.595–612.

[18] Delbaen, F.,1998. Coherent risk measures on general probability spaces. Working paper, ETH Z¨urich. http://www.math.ethz.ch/∼delbaen/.

[19] Dhaene, J. et al., 2006. Risk measures and comonotonicity: A review. Stochastic Models, 22(4), pp.573–606.

[20] Duffie, D. & Pan, J., 1997. An overview of value at risk. Journal of Derivatives, 4(3), pp.7–49.

[21] Embrechts, P. et al., 2014. An Academic Response to Basel 3.5. Risks, 2(1), pp.25–48.

[22] Embrechts, P., Puccetti, G. & Rüschendorf, L., 2013. Model uncertainty and VaR aggregation. Journal of Banking and Finance, 37(8), pp.2750–2764.

[23] Emmer, S., Kratz, M. & Tasche, D., 2015. What is the best risk measure in practice? A comparison of standard measures. Journal of Risk, 18(2), pp.31–60.

[24] Fahad, A. et al., 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing, 2(3), pp.267–279.

[25] Froot, K. & Stein, J., 1998. Risk management, capital budgeting, and capital structure policy for financial institutions: An integrated approach. Journal of Financial Economics, 47(1), pp.55–82.

[26] Gabrel, V., Murat, C. & Thiele, A., 2013. Recent advances in robust optimization: An overview. European Journal of Operational Research, 235(3), pp.471–483.

[27] Goovaerts, Marc J., Kaas, Rob & Laeven, Roger J.A., 2011. Worst case risk measurement: Back to the future? Insurance Mathematics and Economics, 49(3), pp.380–392.

[28] Harmantzis, F.C., Miao, L. & Chien, Y., 2006. Empirical study of value-at-risk and expected shortfall models with heavy tails. The Journal of Risk Finance, 7(2), pp.117–135.

[29] Hendricks, D., 1996. Evaluation of value-at-risk models using historical data. Federal Reserve Bank of New York economic policy review, 2(1), pp.39–69.

[30] Irpino, A. & Verde, R., 2006. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. Data science and classification, IFCS 2006.

[31] Irpino, A. & Verde R., 2008. Dynamic clustering of interval data using a Wasserstein-based distance. Pattern Recognition Letters, 29(11), pp.1648–1658.

[32] Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), pp.651–666.

[33] Jain, A., Murty, M. & Flynn, P., 1999. Data clustering: a review. ACM Computing Surveys (CSUR), 31(3), pp.264–323.

[34] Jorion, P., Value at Risk: The new benchmark for managing financial risk, McGraw-Hill Companies.

[35] Klir, G.J., 1987. Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like? Fuzzy Sets and Systems, 24(2), pp.141–160.

[36] Krätschmer, V., Schied, A. & Zähle, H., 2011. Qualitative and infinitesimal robustness of tail-dependent statistical functionals. Journal of Multivariate Analysis, 103(1), pp.35–47.

[37] Krätschmer, V., Schied, A. & Zähle, H., 2014. Comparative and qualitative robustness for law-invariant risk measures. Finance and Stochastics, 18(2), pp.271–295.

[38] Kratz, M., Lok, Y. & McNeil, A., 2018. Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. Journal of Banking and Finance, 88, pp.393–407.

[39] Liu, W., 2008. Analysis of risk measures and multi-dimensional risk dependence, pp.ProQuest Dissertations and Theses.

[40] Markowitz, H., 1952. Portfolio selection. Journal of Finance, 7(1), pp.77–91.

[41] Mohajerin Esfahani, P. & Kuhn, D., 2018. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Mathematical Programming, 171(1), pp.115–166.

[42] Matthias, F., Thorsten, M. & Marius, P., 2018. A discussion on recent risk measures with application to credit risk: Calculating risk contributions and identifying risk concentrations. Risks, 6(4), 142–170.

[43] Pandit, S. & Gupta, S., 2011. A comparative study on distance measuring approaches for clustering. International Journal of Research in Computer Science, 2011(25), pp.29–31.

[44] Pflug, G. & Wozabal, D., 2007. Ambiguity in portfolio selection. Quantitative Finance, 7(4), pp.435–442.

[45] Pichler, A., 2013. Evaluations of risk measures for different probability measures. SIAM Journal on Optimization, 23(1), pp.530–551.

[46] Rüdiger, K. et al., 2016. The Wasserstein metric and robustness in risk management. Risks, 4(3), pp. 32–46.

[47] Stanislav, P., 2000. Probabilistic constrained optimization: Methodology and applications, Boston, MA: Springer US.

[48] Shumway, R.H., 2003. Time-frequency clustering and discriminant analysis. Statistics and Probability Letters, 63(3), pp.307–314.

[49] Shushang, Z. & Fukushima, M., 2009. Worst-case conditional Value-at-Risk with application to robust portfolio management. Operations Research, 57(5), pp.1155–1168.

[50] Tasche, D., 2002. Expected Shortfall and beyond. Journal of Banking and Finance, 26(7), pp.1519–1533.

[51] Warren Liao, T., 2005. Clustering of time series data - A survey. Pattern Recognition, 38(11), pp.1857–1874.

[52] Wang, X., Smith, K. & Hyndman, R., 2006. Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery, 13(3), pp.335–364.

[53] Yamai, Y. & Yoshiba, T., 2005. Value-at-risk versus expected shortfall: A practical perspective. Journal of Banking and Finance, 29(4), pp.997–1015.

[54] Zhang, T., Ramakrishnan, R. & Livny, M., 1996. An efficient data clustering method for very large databases. SIGMOD Record (ACM Special Interest Group on Management of Data), 25(2), pp.103–114.