Estimation of cardinality constrained portfolio efficiency via segmented DEA

Zhongbao Zhou
School of Business Administration, Hunan University, Changsha 410082, P. R. China, z.b.zhou@hnu.edu.cn
Qianying Jin
School of Business Administration, Hunan University, Changsha 410082, P. R. China, qianyingjin@hnu.edu.cn
Helu Xiao
School of Business Administration, Hunan University, Changsha 410082, P. R. China, xiaohelu1986@163.com
Qian Wu
School of Business Administration, Hunan University, Changsha 410082, P. R. China, wq10082@163.com
Wenbin Liu
KBS, University of Kent, Canterbury, Kent, CT2 7NZ, UK, w.b.liu@kent.ac.uk

The cardinality constrained portfolio selection problem arises due to the empirical findings that investors tend to hold limited number of assets. Yet the lack of efficient frontier of cardinality constrained portfolio investments makes the performance evaluation of this problem a long-standing challenge. Classic Data Envelopment Analysis (DEA) models have been justified valid in evaluating and ranking portfolio performance. Unfortunately, when it comes to the cardinality constrained portfolio selection problem, the DEA models fail to approximate the portfolio efficiency (PE) since the real frontier is discontinuous and not concave. To solve this problem, we propose a segmented DEA approach based on data segment points. A searching algorithm is introduced to approach the real segment points and proved to be valid. In each segment, the frontier is continuous and concave; hence, classic DEA models can be applied to evaluate the PE. The simulation results further indicate that the segmented DEA approach proposed in this paper is effective and practical in evaluating the cardinality constrained portfolio performance.

Key words: cardinality constrained portfolio selections; performance evaluation; data envelopment analysis

1. Introduction

Portfolio performance evaluation has always been a hot issue from both academic and practical viewpoints. Besides the most well-known performance measures, the Treynor index (Treynor, 1965), the Sharpe index (Sharpe, 1966) and the Jensen index (Jensen, 1968) which are still in use today, portfolio frontier approach is the most important idea in portfolio performance evaluation. Since the mean-variance (MV) framework proposed by Markowitz (1952) laid a solid foundation for frontier approach, abundant of researches have been done to extend the idea to fit in the real investment situation. One important assumption is that, according to Markowitz's classical theory, investors construct their portfolios with all assets available in the market. However, extensive empirical literatures show that many investors prefer to limit the number of assets in their portfolio (Goetzman and Kumar, 2008; Gubaydullina and Spiwoks, 2009). Such a gap between the theory and reality motivates abundant of researchers to study the problem defined as cardinality constrained mean-variance (CCMV) portfolio selection problem. When it comes to assessing the portfolio performance, the CCMV frontier is required as the portfolio frontier approach is realized by comparing some distances to the efficient frontier.

The CCMV portfolio selection problem is a special case of cardinality constrained quadratic optimization (CCQO) problems, which has been proved to be, in general, NP-hard (see Welch, 1982). Since Chang et al. (2000) extended the standard model to include cardinality constraints, there is plenty of work on solving the CCMV portfolio selection problem and calculating the frontier. The main solution schemes in the existing literatures can be classified into either analytically approximating or heuristically solving the exact model. The first method can be further divided into two kinds, one is based on tackling the mixed-integer quadratic program reformation and the other is by applying various relaxations and bounding techniques (Syam, 1998; Mansini and Speranza, 1999; Kellerer et al., 2000; Corazza and Favaretto, 2007; Bonami and Lejeune, 2009). For example, Li et al. (2006) put up an exact solution algorithm which obtains an optimal lot solution to CCMV formulation with concave transaction costs. Specifically, they propose a convergent Lagrangian and contour-domain cut method for discrete-feature constrained portfolio selection problems. In the paper of Shaw et al. (2008), a dedicated Lagrangian relaxation method is developed, the approach is able to take advantage of the special structure of the objective function. Bertsimas and Shioda (2009) propose branch-and-bound based algorithms which also take advantage of the special structure of cardinality constrained quadratic optimization problems. And they develop an exact solution scheme by using a convex relaxation. Gao and Li (2013) propose to modify the objective function using some relaxations. In particular, the analytical solutions to these relaxed cardinality constrained problems are all derived. Recently, Zheng et al. (2014) approximate the cardinality function by a piecewise-linear DC function and solve the cardinality constrained convex program directly. Tackling the CCMV problem exactly is of great computational difficulty, many metaheuristics are then developed for this problem. Heuristic methods are typically metaheuristic based on tabu search, genetic algorithms and simulated annealing. Chang et al. (2000) apply three heuristic algorithms based upon genetic algorithms, tabu search and simulated annealing to find the CCMV frontier. Following the work of Chang et al. (2000), heuristic methods for CCMV portfolio selection problems have been studied by many other authors, a detailed literature review can be found in Woodside-Oriakhi et al. (2011).

It can be seen from the literature review that great efforts have been made to solve the CCMV portfolio selection problem. However, the methods developed in the literatures can still be complex and time-consuming. Moreover, the existing methods are very likely to be inapplicable if the assets number is relatively large. The complexity of calculating the exact CCMV frontier makes performance evaluation a remaining challenge.

The models developed along portfolio frontier approach are referred to as diversification models (or nonlinear DEA models) and traditional Data Envelopment Analysis (DEA) models. With the property of diversification in constructing portfolios, diversification models are widely discussed. Morey and Morey (1999) present a quadratic constrained nonlinear DEA model which avoids the limitations of the earlier indices used to identify mutual fund rankings. Briec et al. (2004) extend the work of Morey and Morey (1999) in several ways, especially by introducing a more general efficiency measure. Under the diversification DEA framework, Branda (2013) introduced the diversification-consistent DEA model referring to portfolios with limited number of assets which is known as cardinality constraint here, although the explicit solutions were not discussed. Branda (2015) proposes new diversification-consistent DEA models which can identify weakly, semi-strongly and strongly Pareto efficient portfolios. Furthermore, the diversification models are extended into the non-convex space concerning higher orders (Kerstens et al., 2011, Nalpas et al., 2016). It is natural to apply diversification models in performance evaluation of CCMV problem. However, the application of diversification models is limited due to its complexity and massive computational work. Murthi et al. (1997) first propose DEA as a measure of performance. Subsequently, McMullen and Strong (1998), Galagedera and Silvapulle (2002), Daraio and Simar (2006) extend the applications of DEA models. While taking market frictions and bounds into account, DEA models have been widely applied to evaluate portfolio performances. Liu et al. (2015) systematically investigate the theoretical justifications of applying DEA in estimating portfolio performance under the assumption of convexity. For example, they have shown that DEA can take into account sufficient diversification, therefore produces reliable evaluations by showing that the DEA frontier can converge to the portfolio frontier when there exist adequate data.

Naturally, we will intend to apply the DEA methods to evaluate the portfolio performance of CCMV problem. However, the CCMV frontier is non-concave which may lower the evaluation accuracy if DEA models are applied directly. Nevertheless, studies about the structure of CCMV portfolio optimization problem conclusively show that the CCMV frontier consists of segmented frontiers which share some common properties with classical MV frontier. For example, Chang et al. (2000) shows that each segment of the CCMV frontier between two adjacent segment points is both continuous and concave, although it is still not clear how to identify these segmentations in realistic time. Thus, the performance evaluation of CCMV problem should be carried out in a correct segmentation of the real CCMV frontier. The frontier will be continuous and concave in each segment, and then we can apply classical DEA methods. Hence, the key issue here is to numerically identify the segment points of the CCMV frontier effectively.

The paper is organized as follows. After introduction, the definition of portfolio efficiency (PE) under the circumstance of cardinality constraints is presented in Section 2. In Section 3, a searching algorithm for data segment points is introduced. In addition, the convergence property of the algorithm is studied, which indicates that the data segment points can reliably approximate the real segment points with sufficient data. Then, the DEA models used to estimate the PE are presented. In Section 4, we numerically test our searching algorithm under different problem formulations. We conclude our paper in Section 5.

2. Portfolio Efficiency in cardinality constrained mean-variance framework

2.1 Problem formulation

Suppose there are n assets available in the market. The expected return vector and covariance matrix are given by $r = (r_1, \ldots, r_n)'$ and $G = \{\sigma_{ij}\}_{i,j=1}^{n}$, respectively, where σ_{ij} denotes the covariance between asset i and j. The covariance matrix G is assumed to be positive definite. Let $x = (x_1, \ldots, x_n)' \in \Omega$ be the portfolio weights invested in n risky assets such that $x \in R^n$ and $\sum_{i=1}^{n} x_i = 1$. Ω is the feasible set of portfolio weights. Let K be the desired number of risky assets in constructing portfolios.

Then the CCMV optimization problem can be formulated as follows:

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} x_i \sigma_{ij} x_j \qquad (1a)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} r_i x_i \geq \bar{r} \qquad (1b)$$

$$\sum_{i=1}^{n} x_i = 1 \qquad (1c)$$

$$\sum_{i=1}^{n} \left|\operatorname{sign}(x_i)\right| = K \qquad (1d)$$

where sign function $\operatorname{sign}(\cdot)$ is defined as

$$\operatorname{sign}(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{if } a = 0 \\ -1 & \text{if } a < 0 \end{cases} \quad (2)$$

The total variance associated with the portfolio is optimized to minimum by the objective function $(1a)$ whilst $\bar{r}$ sets a target portfolio return level in $(1b)$. The equation $(1c)$ ensures that all wealth has been invested. Cardinality constraints are presented by $(1d)$, which requires the number of assets in a portfolio is limited to K.

2.2 PE definition

Other than the traditional performance indexes, the frontier-based PE is also an important idea in measuring portfolio performance (Morey and Morey, 1999; Joro and Na, 2006; Glawischnig and Sommersguter-Reichmann, 2010; Brandouy et al, 2015). Following the idea of portfolio frontier approach, PE of a portfolio is defined by a relative distance to the frontier. In a case that the frontier is continuous and concave, different portfolio efficiencies can be defined by using different distances. Suppose there is a sample of m portfolios to be evaluated. For portfolio j ($j = 1, \ldots, m$), assume that $y_j = (y_{1j}, y_{2j}, \ldots, y_{nj})$ represents the portfolio weight vector, then the expected return and its variance are $E(y_j)$ and $V(y_j)$, respectively. As shown in Figure 1, the variance is in accordance with the abscissa, and the ordinate represents the expected return. Let $A_1(E(y_1), V(y_1))$ denotes a portfolio under evaluation. $B_1(E(x_1^*), V(x_1^*))$, $B_2(E(x_2^*), V(x_2^*))$, $B_3(E(x_3^*), V(x_3^*))$ are reference points, that is the optimal portfolios calculated by using return-oriented, risk-oriented and non-oriented measures, respectively (Briec and Kerstens, 2009).

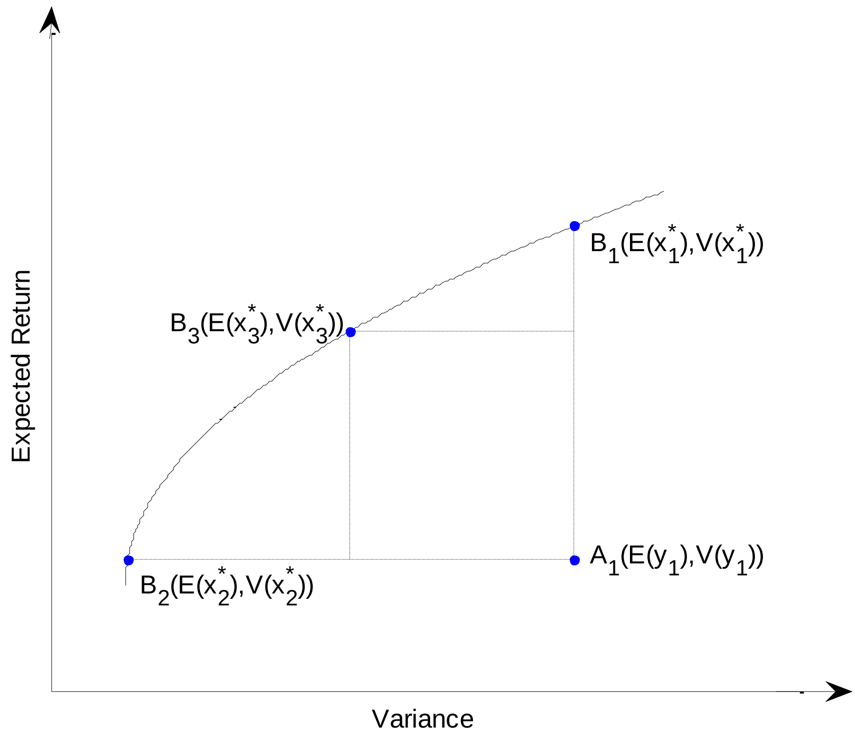


Figure 1. PE definition of classic MV frontier.

Thus, by using different distances, return-oriented, risk-oriented and non-oriented portfolio efficiencies can be defined, respectively.

$$PE_E = \frac{E(y_0)}{E(x^*)},\ \ PE_V = \frac{V(x^*)}{V(y_0)},\ \ PE_N = \frac{1-(V(y_0)-V(x^*))/V(y_0)}{1+(E(x^*)-E(y_0))/E(y_0)} \quad (3)$$

To be more specific, with given directional distance function (DDF) $g = (-g_V, g_E)$, different oriented-PE of portfolio y_0 is calculated by the following model (Briec et al., 2004).

$$\begin{aligned} & \max \delta \\ & \text{s.t.}\quad E(y_0) + \delta g_E \le \sum_{i=1}^{n} x_i r_i \\ & \qquad V(y_0) + \delta g_V \ge \sum_{i=1}^{n}\sum_{j=1}^{n} x_i \sigma_{ij} x_j \ (3) \\ & \qquad \sum_{i=1}^{n} x_i = 1 \\ & \qquad x_i \ge 0,\ i = 1, \dots, n \end{aligned}$$

If $g_V = 0$ and $g_E = E(y_0)$, the return-oriented PE is defined as $PE_E = \frac{E(y_0)}{E(x^*)} = \frac{1}{1+\delta}$. If $g_V = -V(y_0)$ and $g_E = 0$, the risk-oriented PE is defined as $PE_V = \frac{V(x^*)}{V(y_0)} = 1-\delta$. If $g_V = -V(y_0)$ and $g_E = E(y_0)$, then the non-oriented PE is defined as $PE_N = \frac{1-(V(y_0)-V(x^*))/V(y_0)}{1+(E(x^*)-E(y_0))/E(y_0)} = \frac{1-\delta}{1+\delta}$.

One aspect of CCMV portfolio optimization problem that appears to have received extensive attention in the literatures is the fact that the CCMV efficient frontier is markedly different from the classical MV frontier. Taken as a whole, the CCMV efficient frontier may be discontinuous, which may lead to the absence of reference points in certain orientation. Hence in the CCMV case, it is not always possible to define PE under every orientation, and this will depend on the geometrical feature of the CCMV frontier. For the case that the frontier is continuous (see Figure 2), it is possible to define both return-oriented and risk-oriented portfolio efficiencies.

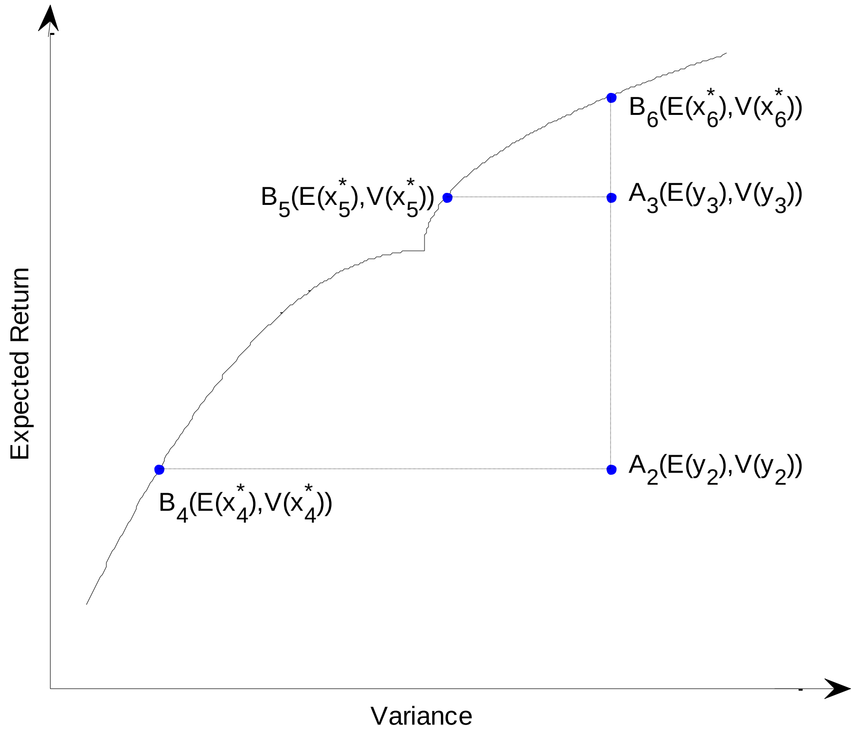


Figure 2. PE definition of continuous CCMV frontier.

As shown in Figure 2, $B_6\left(E(x_6^*),V(x_6^*)\right)$ is a return-oriented reference point for both $A_2\left(E(y_2),V(y_2)\right)$ and $A_3\left(E(y_3),V(y_3)\right)$, while $B_4\left(E(x_4^*),V(x_4^*)\right)$ and $B_5\left(E(x_5^*),V(x_5^*)\right)$ are risk-oriented reference points for $A_2\left(E(y_2),V(y_2)\right)$ and $A_3\left(E(y_3),V(y_3)\right)$, respectively. Hence, for points like A_2 and A_3 which are under a continuous CCMV frontier, different portfolio efficiencies in (3) can all be defined.

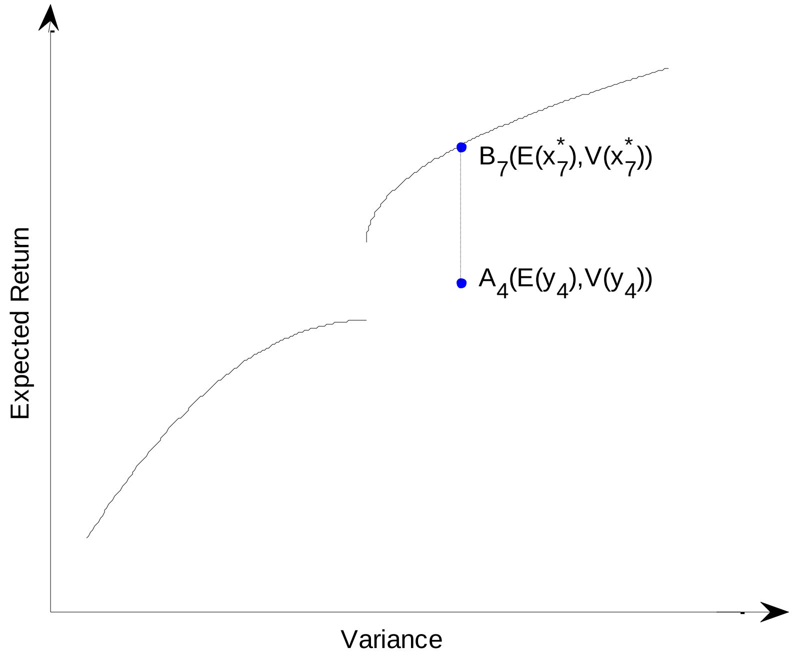


Figure 3. PE definition of discontinuous CCMV frontier: returned-oriented.

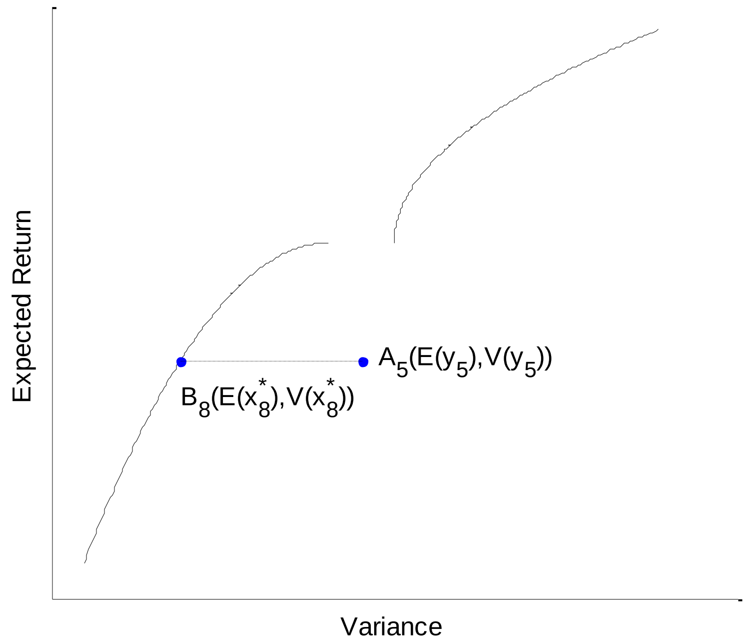


Figure 4. PE definition of discontinuous CCMV frontier: risk-oriented.

It will make a significant difference for discontinuous frontiers. As shown in Figure 3, $A_4\left(E(y_4),V(y_4)\right)$ denotes the portfolio under evaluation, $B_7\left(E(x_7^*),V(x_7^*)\right)$ denotes a reference point sharing the same variance with A_4. Since there is a jump on the CCMV frontier, point with the same return comparing to point A_4 is not exist. Hence, only the return-oriented PE is available for A_4. On the contrary, in Figure 4, $A_5\left(E(y_5),V(y_5)\right)$ denotes the portfolio under evaluation, $B_8\left(E(x_8^*),V(x_8^*)\right)$ denotes the reference point. The risk here is preferred as an orientation since no points on the efficient frontier share the same risk with A_5. This situation appears when considering the lower and upper bounds of portfolio weights.

3. Estimation of PE for CCMV problems via segmented DEA model

In this section, a searching algorithm is introduced to locate the data segment points, which are then proved to approximate the real segment points. Consequently, classic DEA models can be used in each segment. The expected return is considered as a desirable output, while the variance is an undesirable output, we will treat it as an input in this paper. Estimating the mean-standard deviation (M-SD) portfolio frontier may lead to the choice of different DEA models. If no risk-free assets exist, the BCC model (Banker et al., 1984) is a good choice no matter shorting is allowed or not. When there are risk-free assets, NIRS DEA model (Färe and Grosskopf, 1985) can be used if short-sale is not allowed, otherwise the CCR model (Charnes et al., 1978) is used (after shifting the return of risk-free asset). When it comes to the M-V problem, BCC model would always be a good choice to approach the portfolio frontier, since the M-V frontiers are always concave. Concerning the CCMV problem discussed in this paper, we will only discuss the BCC models below.

Definition 1. The CCMV portfolio frontier function is denoted by $r = F(\sigma)$. A point $(\tilde{\sigma}_i, \tilde{r}_i)(\tilde{r}_i = F(\tilde{\sigma}_i))$ $(i = 1,2,\ldots)$ will be defined as a real segment point only when it satisfies the conditions given below.

1) The CCMV frontier $F(\sigma)$ defined in the segment $[\tilde{\sigma}_i, \tilde{\sigma}_{i+1}](i = 1,2,\ldots)$ is smooth.

2) A point is either of the two types below:

Type 1: a discontinuous point that satisfies $\lim_{\sigma \to \tilde{\sigma}_i^-} F(\sigma_i) \neq \lim_{\sigma \to \tilde{\sigma}_i^+} F(\sigma_i)$;

Type 2: a continuous point but with inequality between the derivatives, that is $\lim_{\sigma \to \tilde{\sigma}_i^-} F(\sigma_i) = \lim_{\sigma \to \tilde{\sigma}_i^+} F(\sigma_i)$ but $F'(\sigma_i^+) > F'(\sigma_i^-)$.

3.1 Searching algorithm for data segment points

The algorithm for searching data segment points is consisting of two parts. First, it takes three steps to get the outermost layer of the original sample points. After that, the searching process is introduced to locate the data segment points. The detail of our searching algorithm is shown as follows:

(1) Sort the sample points by variance and obtain a sequence (σ_a, r_a) $(a = 1,2,...,n)$, where σ_a is in a not descending order.

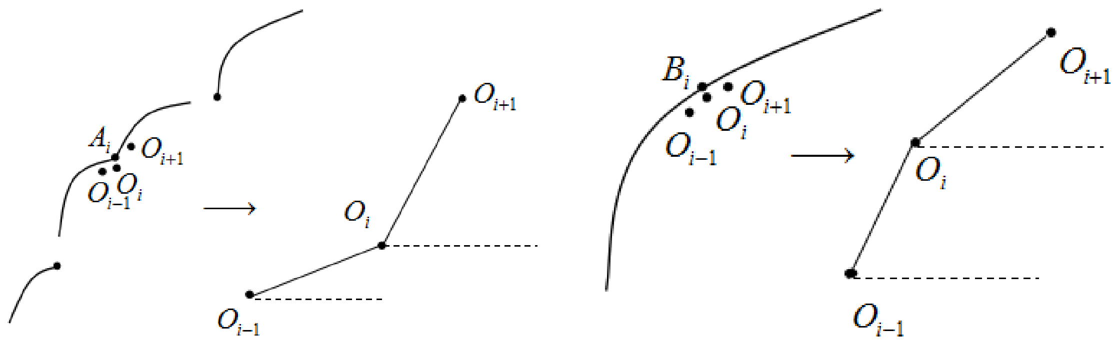
(2) Let $\hat{\sigma}_b = \sigma_a$ and $\hat{r}_b = \max_{a1 \le k \le a2}(r_k)$ when the situation $\sigma_{a1-1} \neq \sigma_{a1} = \sigma_{a1+1} = \cdots = \sigma_{a2} \neq \sigma_{a2+1}$ appears, otherwise simply record $\hat{\sigma}_b = \sigma_a$ and $\hat{r}_b = r_a$. With this process, a new sequence $(\hat{\sigma}_b, \hat{r}_b)$ $(b = 1,2,...,\hat{n})$ is obtained, where $\hat{\sigma}_b$ is strictly monotone increasing ($\hat{\sigma}_b < \hat{\sigma}_{b+1}, b = 1,2,...,\hat{n}-1$).

(3) Let $\widehat{\sigma}_i = \hat{\sigma}_b$ and $\widehat{r}_i = \hat{r}_b$ if $\hat{r}_b \le \hat{r}_{b+1}$, otherwise sample $(\hat{\sigma}_{b+1}, \hat{r}_{b+1})$ is abandoned. In the new sequence $(\widehat{\sigma}_i, \widehat{r}_i)$ $(i = 1,2,...,\widehat{n})$, $\widehat{r}_i$ is monotone increasing ($\widehat{r}_i \le \widehat{r}_{i+1} (i = 1,2,\ldots,\widehat{n}-1)$) while $\widehat{\sigma}_i$ is strictly monotone increasing ($\widehat{\sigma}_i < \widehat{\sigma}_{i+1}, i = 1,2,...,\widehat{n}-1$).

(4) For any given sample point $(\widehat{\sigma}_i, \widehat{r}_i)$ $(i = 2,...,\widehat{n}-1)$, define two slopes: $k_{i-j_1} = \dfrac{\widehat{r}_{i-j_1} - \widehat{r}_i}{\widehat{\sigma}_{i-j_1} - \widehat{\sigma}_i}$ $(j_1 = 1,2,\ldots,L_1 - 1)$ and $k_{i+j_2} = \dfrac{\widehat{r}_{i+j_2} - \widehat{r}_i}{\widehat{\sigma}_{i+j_2} - \widehat{\sigma}_i}$ $(j_2 = 1,2,\ldots)$, respectively, where $(\widehat{\sigma}_{i-L_1}, \widehat{r}_{i-L_1})$ is a known data segment point. If there exist $j_{01} \in \{ j_1 | j_1 = 1,2,\ldots,L_1 - 1\}$

and $j_{02} \in \{ j_2 | j_2 = 1,2,\ldots\}$ which make the inequality $k_{i-j_{01}} < k_{i+j_{02}}$ satisfied, then $(\hat{\sigma}_i, \hat{r}_i)$ is identified as a data segment point.

The searching algorithm based on the discrete sample points is expected to locate the data segment points, consequently, DEA models can be applied in each segment. Below we will illustrate the main idea behind our algorithm for searching data segment points. Figure 5 shows segmented frontiers and a data point O_i to be examined. Note that the sample points we discussed below are the sample points that have been processed by step (1) to (3).



(a) data segment points (b) normal data points

Figure 5. The principle of the searching algorithm for data segment points.

Let O_{i-1} and O_{i+1} be the sample points adjacent to O_i in both sides, k_{O_i-} be the slope of line $O_{i-1}O_i$, and k_{O_i+} be the slope of line O_iO_{i+1}. According to the property of real CCMV frontier, for a real jump point, the slopes of link lines of its adjacent points will likely to increase, otherwise, for any other point on the frontier, the slopes of link lines of its adjacent points are likely to decrease. Consequently, if $k_{O_i-} < k_{O_i+}$ is satisfied as shown in Figure 5(a), then point O_i is defined as a data segment point, otherwise if $k_{O_i-} \geq k_{O_i+}$, then point O_i is referred to as a normal data point (see Figure 5(b)).

3.2 Theoretical foundation of segment point search algorithm: convergence property

Assumption. Suppose there exists a probability density function $p(x)$ of $x \in \Omega$ satisfying

$\forall x^a \in \Omega$, and there exists a set $S(x^a, \xi) = U(x^a, \xi) \cap \Omega$ such that $\int_{S(x^a, \xi)} p(x)dx > 0$, where $U(x^a, \xi)$ is any neighborhood of x^a.

Theorem 1. Let $r = F(\sigma)$ be the portfolio frontier without risk-free assets and $r^* = F_n^*(\sigma)$ be the BCC-DEA frontier with n portfolio samples. Then $F_n^*(\sigma)$ converges to $F(\sigma)$ in probability when $n \to +\infty$.

Proof. See in Liu et al (2015).

Theorem 2. Let $A = (\tilde{\sigma}_i, \tilde{r}_i)$ be a real segment point of CCMV efficient frontier. In any neighborhood of A, that is $U(A, \varepsilon)(\forall \varepsilon > 0)$, for n large enough, there are always data segment points $O^n = (\hat{\sigma}^n, \hat{r}^n) \in U(A, \varepsilon)$, which converge to the real segment point A. In addition, for all sample points $(\hat{\sigma}_j^n, \hat{r}_j^n) \in U(A, \varepsilon)(j = 1, 2, \cdots)$, there have $\hat{\sigma}^n \geq \hat{\sigma}_j^n$ and $\hat{r}^n \geq \hat{r}_j^n$.

Proof. See in Appendix.

The above theorems essentially indicate that the data segment points found by our searching algorithm will approximate corresponding real segment points. Therefore, it is reasonable to use the data segment points to segment the CCMV frontier so as to apply suitable DEA models.

3.3 DEA models to estimate PE

The real segment points can be properly approximated by the data segment points which are located by our searching algorithm. Hence, the CCMV portfolio frontier can be decomposed into several continuous and concave frontiers. The DEA models can be used to estimate the PE in each segment. It is worth noting that there may be more data segment points than real segment points, but this will not cause mis-estimation although it will increase computational work.

Suppose there are totally m portfolios. Between adjacent data segment points l and $l+1$ $(l = 1, 2 \ldots)$, there are m^l portfolios under evaluation, where $\sum_l m^l = m$. For portfolio j

$(j=1,\ldots,m^l)$, assume that $y_j^l=(y_{1j}^l,y_{2j}^l,\ldots,y_{nj}^l)$ represents the portfolio weight vector, then $E(y_j^l)=\sum_{i=1}^{n}y_{ij}^l r_i$ and $V(y_j^l)=\sum_{i=1}^{n}\sum_{k=1}^{n}y_{ij}^l\sigma_{ik}y_{ij}^l$ are defined as expected return and variance, respectively. In each segment, the following BCC models are properly selected to approximate the efficiency, DMU 0 is a sample point under evaluation.

a) BCC model with risk-oriented measure:

$$
\begin{aligned}
&\min\ \theta\\
&\text{s.t.}\ \sum_{j=1}^{m^l}\lambda_j^l V(y_j^l)\le \theta V(y_0^l)\\
&\qquad \sum_{j=1}^{m^l}\lambda_j^l E(y_j^l)\ge E(y_0^l) \qquad (4)\\
&\qquad \sum_{j=1}^{m^l}\lambda_j^l=1\\
&\qquad \lambda_j^l\ge 0,\ j=1,\ldots,m^l
\end{aligned}
$$

b) BCC model with return-oriented measure:

$$
\begin{aligned}
&\max\ \varphi\\
&\text{s.t.}\ \sum_{j=1}^{m^l}\lambda_j^l V(y_j^l)\le V(y_0^l)\\
&\qquad \sum_{j=1}^{m^l}\lambda_j^l E(y_j^l)\ge \varphi E(y_0^l) \qquad (5)\\
&\qquad \sum_{j=1}^{m^l}\lambda_j^l=1\\
&\qquad \lambda_j^l\ge 0,\ j=1,\ldots,m^l
\end{aligned}
$$

It is worth noting that the BCC model used for CCMV portfolio problem is just a classic DEA model, the cardinality constraints are reflected in the data that we used.

4. Simulation

To verify the validity of the solution schemes proposed above, we construct some investment situations by using the historical data from GSMAR database provided by the GTA Information Technology Corporation. In particular, we choose five stocks and use their daily return data between

January 2015 and August 2015 to estimate their mean and variance. The statistical properties are shown in Table 1.

Table 1. Statistical properties of the stock pool.

	1	2	3	4	5
Mean	1.3403	1.3124	1.1230	0.9230	1.3872
Covariance	1	2	3	4	5
1	1.9907	0.0146	0.2617	-0.3112	0.4015
2	0.0146	1.5835	-0.2274	0.0187	0.4370
3	0.2617	-0.2274	1.1592	-0.3382	-0.0763
4	-0.3112	0.0187	-0.3382	1.0889	0.2265
5	0.4015	0.4370	-0.0763	0.2265	2.0472

All the problems are solved under the Windows 7 platform (2.0 G CPU, 4 GB RAM). In particular, the CCMV problem is solved by YALMIP (Lofberg, 2004) and the DEA models are calculated by Matlab 2013a. We first produce investment weights in a discrete uniform distribution and use them to construct sample points with known expected returns and variances. Then, the PE of every sample point is derived by comparing its distance to the optimal point on the frontier, and DEA scores are calculated by applying DEA models in each segment. We compare the evaluation results of our approach with that of the YALMIP solver. To test the applicability and superiority of the segmented DEA approach proposed, the correlation coefficients of PEs and DEA scores as well as the correlation coefficients of their ranks are compared under different sample sizes. Furthermore, the CPU time that different approaches consumed is record as well. To be more specific, the time consumption of the segmented DEA approach contains two parts, one is the time used to locate the data segment points via the searching algorithm, the other is the time used to estimate portfolio efficiencies via the segmented DEA approach.

4.1 CCMV model without no-shorting constraints

Assume investors choose three out of five stocks to construct portfolios. We generate weight vectors $x = (x_1, \dots, x_5)' \in \Omega$ under different sample sizes, where $\Omega = \left\{ (x_1, \dots, x_5) \middle| \sum_{i=1}^{5} x_i = 1, \sum_{i=1}^{n} |\mathrm{sign}(x_i)| = 3, i = 1, \dots, 5 \right\}$ and the sample size m is 10, 100, 500 and 1000, respectively.

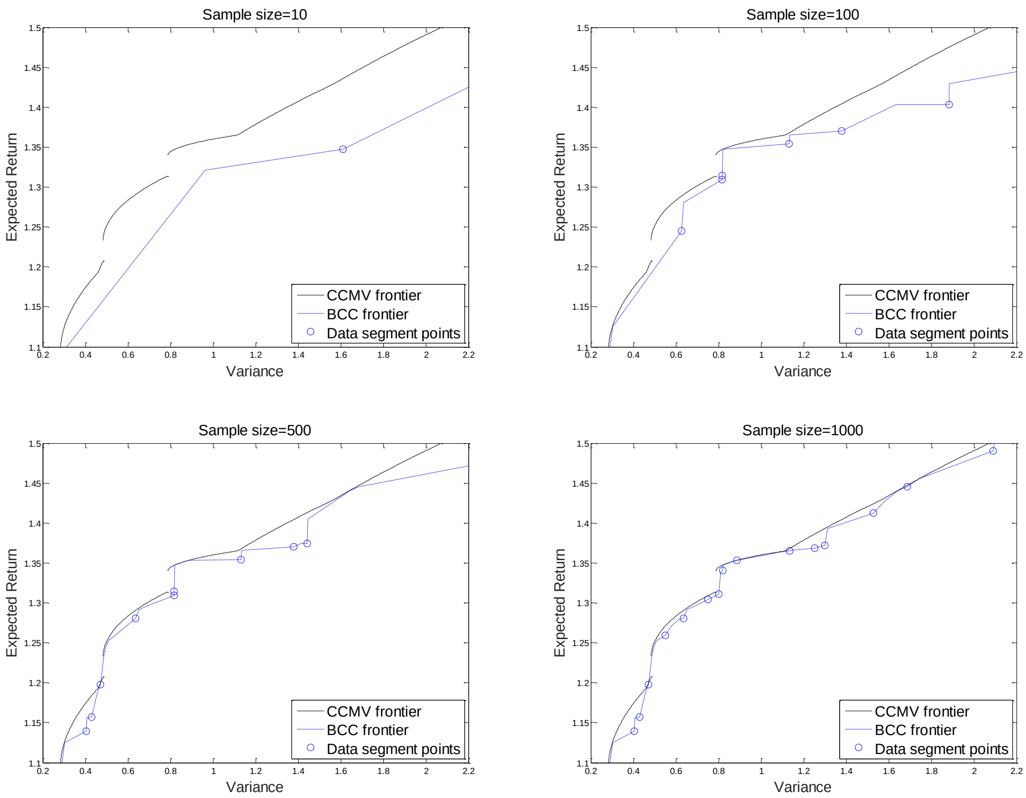


Figure 6. Data segment points location with different sample sizes.

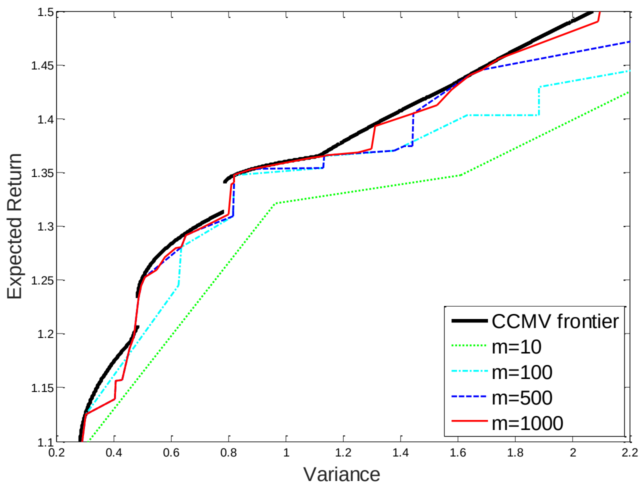


Figure 7. Frontier comparison with different sample sizes

From Figure 6 and 7, it is obvious that when sample size is increasing, the data segment points located by our algorithm are increasing, and the DEA envelopment frontier is approaching the CCMV frontier significantly.

Table 2. Correlation coefficients of efficiencies and ranks with different sample sizes.

sample size	10	100	500	1000
efficiency	0.7749	0.8407	0.8480	0.8711
rank	0.8667	0.9636	0.9273	0.9273

Table 3. Time consumption of YALMIP solver and segment DEA approach.

sample size	10	100	500	1000
YALMIP	3.852	37.345	176.5982	373.7297
DEA	0.2306	2.8665	15.2126	30.1662

In Table 2, the correlation coefficients of scores and ranks verify that the DEA envelopment frontier is approaching the CCMV frontier mathematically. In particular, the approaching methods show good results even with small samples, which is significant for applications in reality. Second, from Table 3, we can conclude that the time YALMIP used to calculate the PE is at least eleven times than that the DEA model used.

4.2 CCMV model with no-shorting constraints

Assume investors choose three out of five stocks to construct portfolios. We generate weight vectors $x=(x_1,\ldots,x_5)'\in\Omega$ under different sample sizes, where $\Omega=\left\{(x_1,\ldots,x_5)\left|\sum_{i=1}^{5}x_i=1,\sum_{i=1}^{n}\left|\operatorname{sign}(x_i)\right|=3,\,0\le x_i\le 1,i=1,\ldots,5\right.\right\}$ and the sample size m is 10, 100, 500 and 1000, respectively. Note that short-selling is not allowed in this case.

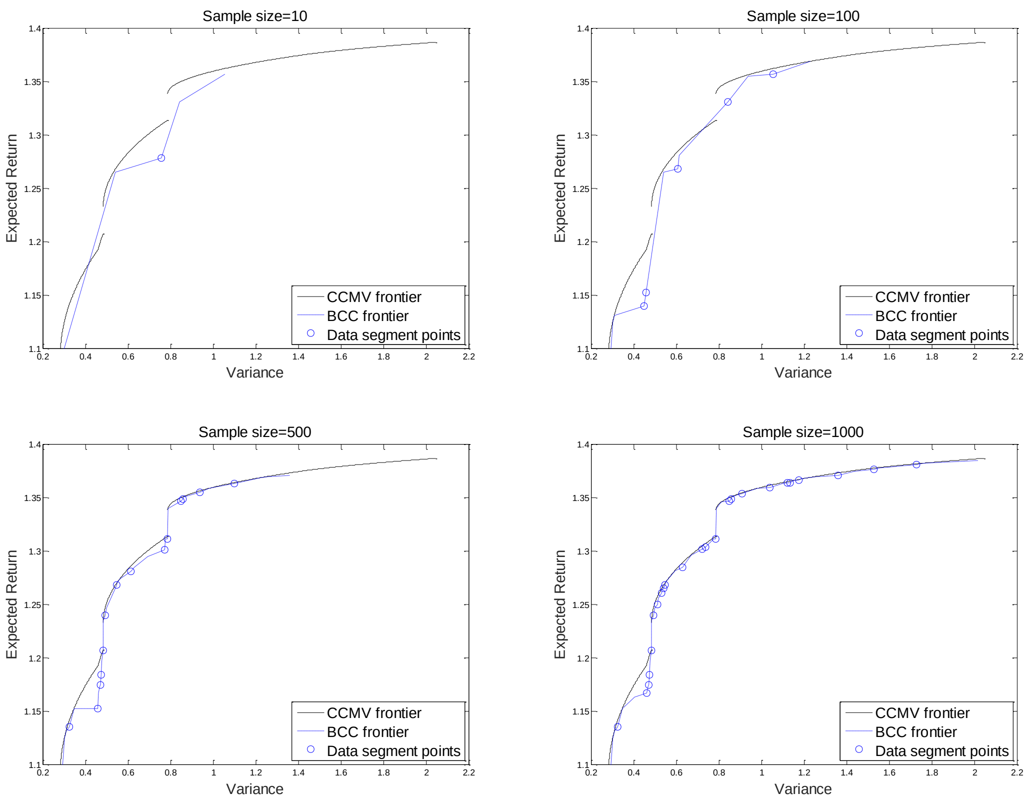


Figure 8. Data segment points location with different sample sizes.

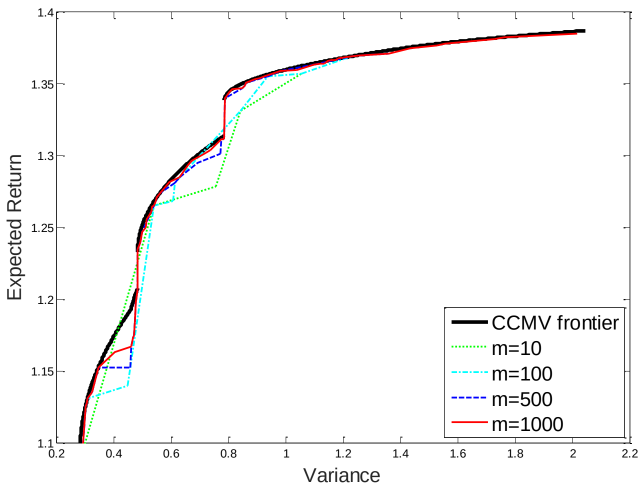


Figure 9. Frontier comparison with different sample sizes.

The results implied by Figure 8-9 are similar to those of Figure 6-7. When sample size is increasing, the data segment points located by our algorithm are increasing, and the DEA envelopment frontier is approaching the CCMV frontier. It implies that the extra constraint of no-shorting has limited effect in applying segmented DEA approach to estimate the portfolio frontier.

Table 4. Correlation coefficients of efficiencies and ranks with different sample sizes.

sample size	10	100	500	1000
efficiency	0.8788	0.8117	0.8601	0.8754
rank	0.8788	0.8182	0.8788	0.8788

Table 5. Time consumption of YALMIP solver and segment DEA approach.

sample size	10	100	500	1000
YALMIP	4.2219	42.0066	212.4776	450.094
DEA	0.2415	2.733	16.2347	33.5632

In Table 4, the correlation coefficients of scores and ranks verify that the DEA envelopment frontier is approaching the CCMV frontier mathematically. What's more, the approaching methods still show good results with small samples. Second, from Table 5, we can conclude that the time YALMIP used to calculate the PE is much less than that the DEA model used. Comparing Table 5 with Table 3, incorporating a no-shorting constraint into the problem causes a significant increase in the time that YALMIP consumed while limited change to the segmented DEA approach. As a NP-hard problem, solving the CCMV portfolio selection problem remains time-consuming, however, the DEA model we applied provides a much easier and quicker way to evaluate the PE.

4.3 CCMV model with increasing asset pool

To evaluate the computational performance of our approach compared with YALMIP solver, we construct test problems with increasing problem dimension n and the cardinality of the portfolio K. Based on the constituent stock of Shanghai Stock Exchange, we choose 648 stocks which have a full observation with a sample size of 247 to form the stock pool and use their daily return data between March 2015 and March 2016 to estimate the mean and covariance matrix. We compare the performance of our approach with the standard YALMIP solver. For the YALMIP solver, we set the stopping criteria of gap ratio as 1E-06 and an upper bound of execution time as 1800 seconds. The computational results are listed below, where the columns "Time", "Suc" record the CPU time in seconds and the number of problems being solved successfully (in total 10 problems).

Table 6. Comparison of DEA approach and YALMIP solver for CCMV problem

		YALMIP		DEA				YALMIP		DEA	
n	K	Time	Suc	Time	Suc	n	K	Time	Suc	Time	Suc
50	10	0.5	10	0.0261	10	50	25	7.8	10	0.0271	10
100	10	0.7	10	0.0265	10	80	40	55.2	10	0.0267	10
200	10	2.4	10	0.0255	10	100	50	562.6	8	0.0277	10
300	10	4.7	10	0.0245	10	140	70	1165.8	4	0.0234	10
400	10	217.5	9	0.0263	10	180	90	1538.2	2	0.0253	10
500	10	5.9	10	0.0240	10	200	100	969.9	5	0.0234	10
540	10	1800.0	0	0.0258	10	240	120	1773.7	1	0.0280	10
580	10	1800.0	0	0.0244	10	280	140	1800.0	0	0.0258	10
600	10	1800.0	0	0.0261	10	300	150	1800.0	0	0.0465	10

We can observe from Table 6 that our segmented DEA approach comes up with solutions for all 10 problems, while the YALMIP solver solves only part of them. That is, the increase of n and K adds the complexity of solving problems in YALMIP solver, while it has limited effect on using segmented DEA approach, since DEA model is a linear program. Furthermore, compared with the YALMIP solver, the segmented DEA approach consumes significantly less CPU time. When n and K increase, one may argue that a larger sample size is needed for the segmented DEA approach to assure the accuracy of PE estimation. It is worth noting that the increase of sample size will only lead to linear growth of the problem complexity and can be realized at a minimal computational cost. Generally speaking, for the CCMV problem, the segmented DEA approach performs much better than YALMIP in the aspect of time consumption.

5. Conclusion

Motivated by the long-standing challenge of the performance evaluation under CCMV framework, we have investigated in this paper an applicable way of applying portfolio frontier approach. In contrast to the existing literatures that have primarily focused on some direct methods to calculate the CCMV frontier, we choose to approximate the CCMV frontier by innovatively applying DEA methods in this paper. First, we define the PE with the consideration of the special geometric properties presented by CCMV frontier. Then, a searching algorithm for data segment points is introduced and proved to be valid for approaching the real segment points. In each segment, DEA models are applied. DEA is a linear modeling tool which makes it computational viable for large scale CCMV portfolio problems. The proposed solution scheme is verified by the simulations.

Our new approach is flexible in allowing adjustment to return function, also allowing restrictions on portfolio weights, and among others (Liu et al., 2015). And similar to the idea of using a polynomial goal programing to solve multi-objective program (Lai, 1991), the higher-degree polynomial forms of the portfolio utility function can be incorporated into our approach by combining polynomial elements with existing inputs/ outputs (e.g., the input is the combination of variance and kurtosis). In a sense, our approach is applicable as long as the portfolio frontier can show convexity after proper segmentation. Of course, we do not pretend that our approach would provide an answer to all objections formulated to standard MV model (for example, to consider multi-horizon portfolio performance). These questions remain to be tackled in the future work.

Reference

[1] Banker R D, Charnes A, Cooper W W. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management science, 30(9): 1078-1092.

[2] Bertsimas D, Shioda R. 2009. Algorithm for cardinality-constrained quadratic optimization. Computational Optimization and Applications, 43(1): 1-22.

[3] Bonami P, Lejeune M A. 2009. An exact solution approach for portfolio optimization problems under stochastic and integer constraints. Operations research, 57(3): 650-670.

[4] Branda M. 2013. Diversification-consistent data envelopment analysis with general deviation measures[J]. European Journal of Operational Research, 226(3): 626-635.

[5] Branda M. 2015. Diversification-consistent data envelopment analysis based on directional-distance measures. Omega, 52: 65-76.

[6] Brandouy O, Kerstens K, Van de Woestyne I. Frontier-based vs. traditional mutual fund ratings: a first backtesting analysis[J]. European Journal of Operational Research, 2015, 242(1): 332-342.

[7] Briec W, Kerstens K, Lesourd J B. Single-period Markowitz portfolio selection, performance gauging, and duality: a variation on the Luenberger shortage function[J]. Journal of Optimization Theory and Applications, 2004, 120(1): 1-27.

[8] Briec W, Kerstens K. 2009. Multi-horizon Markowitz portfolio performance appraisals: A general approach. Omega, 37(1): 50-62.

[9] Chang T J, Meade N, Beasley J E, et al. 2000. Heuristics for cardinality constrained portfolio optimisation. Computers & Operations Research, 27(13): 1271-1302.

[10] Charnes A, Cooper W W, Rhodes E. 1978. Measuring the efficiency of decision making units. European journal of operational research, 2(6): 429-444.

[11] Corazza M, Favaretto D. 2007. On the existence of solutions to the quadratic mixed-integer mean-variance portfolio selection problem. European Journal of Operational Research, 176(3): 1947-1960.

[12] Daraio C, Simar L. 2006. A robust nonparametric approach to evaluate and explain the performance of mutual funds. European Journal of Operational Research, 175(1): 516-542.

[13] Färe R, Grosskopf S. 1985. A nonparametric cost approach to scale efficiency. The Scandinavian Journal of Economics: 594-604.

[14] Galagedera D U A, Silvapulle P. 2002. Australian mutual fund performance appraisal using data envelopment analysis. Managerial Finance, 28(9): 60-73.

[15] Gao J, Li D. 2013. Optimal cardinality constrained portfolio selection. Operations Research, 61(3): 745-761.

[16] Glawischnig M, Sommersguter-Reichmann M. Assessing the performance of alternative investments using non-parametric efficiency measurement approaches: Is it convincing?[J]. Journal of Banking & Finance, 2010, 34(2): 295-303.

[17] Goetzman W N, Kumar A. 2008. Equity portfolio diversification*. Review of Finance, 12(3): 433-463.

[18] Gubaydullina Z, Spiwoks M. 2009. Portfolio diversification: an experimental study. University of Goettingen, Department of Economics.

[19] Jensen M C. 1968. The performance of mutual funds in the period 1945–1964. The Journal of finance, 23(2): 389-416.

[20] Joro T, Na P. 2006. Portfolio performance evaluation in a mean–variance–skewness framework. European Journal of Operational Research, 175(1): 446-461.

[21] Kellerer H, Mansini R, Speranza M G. 2000. Selecting portfolios with fixed costs and minimum transaction lots. Annals of Operations Research, 99(1-4): 287-304.

[22] Kerstens K, Mounir A, Van de Woestyne I. Geometric representation of the mean–variance–skewness portfolio frontier based upon the shortage function[J]. European Journal of Operational Research, 2011, 210(1): 81-94.

[23] Lai T Y. Portfolio selection with skewness: a multiple-objective approach[J]. Review of Quantitative Finance and Accounting, 1991, 1(3): 293-305.

[24] Li D, Sun X, Wang J. 2006. Optimal lot solution to cardinality constrained mean–variance formulation for portfolio selection. Mathematical Finance, 16(1): 83-101.

[25] Liu W, Zhou Z, Liu D, et al. 2015. Estimation of portfolio efficiency via DEA. Omega, 52: 107-118.

[26] Mansini R, Speranza M G. 1999. Heuristic algorithms for the portfolio selection problem with minimum transaction lots. European Journal of Operational Research, 114(2): 219-233.

[27] Lofberg J. 2004. YALMIP: A toolbox for modeling and optimization in MATLAB[C]//Computer Aided Control Systems Design, 2004 IEEE International Symposium on. IEEE: 284-289.

[28] Markowitz H. 1952. Portfolio selection. The journal of finance, 7(1): 77-91.

[29] McMullen P R, Strong R A. 1998. Selection of mutual funds using data envelopment analysis. The Journal of Business and Economic Studies, 4(1): 1.

[30] Morey M R, Morey R C. 1999. Mutual fund performance appraisals: a multi-horizon perspective with endogenous benchmarking. Omega, 27(2): 241-258.

[31] Murthi B P S, Choi Y K, Desai P. 1997. Efficiency of mutual funds and portfolio performance measurement: A non-parametric approach. European Journal of Operational Research, 98(2): 408-418.

[32] Nalpas, N., Simar, L., & Vanhems, A. (2016). Portfolio Selection in a Multi-Input Multi-Output Setting: a Simple Monte-Carlo-FDH Algorithm (No. ISBA Discussion Paper (2016/22)). UCL.

[33] Sharpe W F. 1966. Mutual fund performance. The Journal of business, 39(1): 119-138.

[34] Shaw D X, Liu S, Kopman L. 2008. Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. Optimisation Methods & Software, 23(3): 411-420.

[35] Syam S S. 1998. A dual ascent method for the portfolio selection problem with multiple constraints and linked proposals. European journal of operational research, 108(1): 196-207.

[36] Treynor J L. 1965. How to rate management of investment funds. Harvard business review, 43(1):

63-75.

[37] Welch W J. 1982. Algorithmic complexity: three NP-hard problems in computational statistics. Journal of Statistical Computation and Simulation, 15(1): 17-25.

[38] Woodside-Oriakhi M, Lucas C, Beasley J E. 2011. Heuristic algorithms for the cardinality constrained efficient frontier. European Journal of Operational Research, 213(3): 538-550.

[39] Zheng X, Sun X, Li D, et al. 2014. Successive convex approximations to cardinality-constrained convex programs: a piecewise-linear DC approach. Computational Optimization and Applications, 59(1-2): 379-397.

Appendix

Proof of Theorem 2.

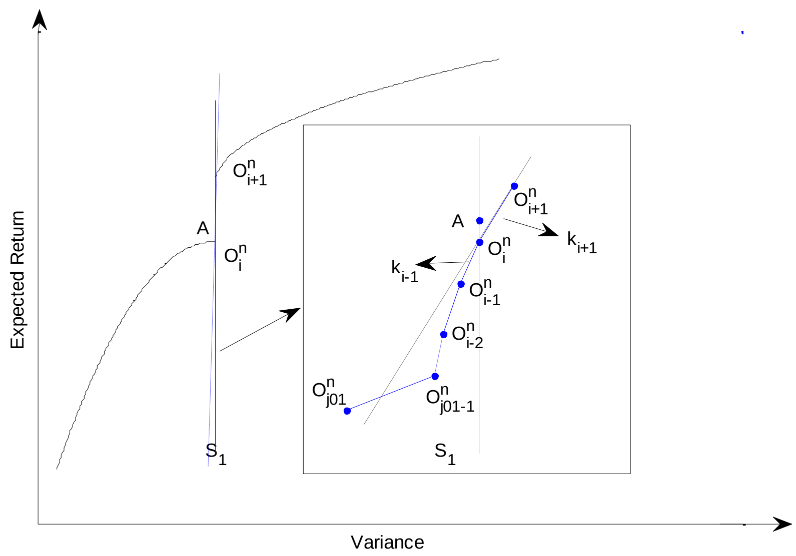


Figure 10. Convergence explanation of Type 1 segment point.

Let $A=(\tilde{\sigma},\tilde{r})$ be a real segment point of Type 1 on the CCMV efficient frontier. Let $k(x)=(\sigma(x),r(x))$ be the risk measure and the expected return measure with the weight vector x.

According to Theorem 1, for any small neighborhood of A, $\int_{k^{-1}(U(A,\varepsilon))\cap\Omega} p(x) > 0$. There exists a sample point $(\hat{\sigma}_i^n, \hat{r}_i^n) \in U(A,\varepsilon)\cap k(\Omega)$ (for n large enough, this is true).

Let: $S_1 = \{(\sigma, r) | \sigma - \hat{\sigma}_i^n \le k_{i+1}\cdot(r - \hat{r}_i^n), 0 \le \sigma \le \hat{\sigma}_i^n, 0 \le r \le \hat{r}_i^n\}$.

If $k_{i-j_1} \ge k_{i+1}$, that is $(\hat{\sigma}_{i-j_1}^n, \hat{r}_{i-j_1}^n) \in S_1$, $(j_1 = 1,2,\ldots)$ (see Figure 10).

Since A is a real segment point, there is a neighborhood in $(S_1)^c \cap k(\Omega_i)$ so that $\int_{k^{-1}((S_1)^c\cap k(\Omega_i))} p(x)dx > \gamma > 0$, where $\Omega_i = \{x | \tilde{\sigma}_{i-1} \le x'Gx \le \tilde{\sigma}_i, \tilde{r}_{i-1} \le r'x \le \tilde{r}_i, I'x = 1, \sum_{i=1}^{n} |sign(x_i)| = K\}$. Thus let T_{j_1} represents the event that the sample point $(\hat{\sigma}_{i-j_1}^n, \hat{r}_{i-j_1}^n)$ $(j_1 = 1,2,\ldots, j_{01}-1)$ is not in the triangular area S_1. Therefore, the probability of T_{j_1} has the property as follows:

$$\Pr(T_{j_1}) \ge \delta > 0 \quad (6)$$

Let T represents the event that not a single sample point $(\hat{\sigma}_{i-j_1}^n, \hat{r}_{i-j_1}^n)$ $(j_1 = 1,2,\ldots)$ of all n portfolio samples is not in the triangular area S_1. Then, the probability of T satisfies the following formula: $\Pr(T) \le (1-\delta)^{j_{01}} \to 0$. Consequently, in probability, there always exists a bounded number j_{01} such that $(\hat{\sigma}_{i-j_{01}}^n, \hat{r}_{i-j_{01}}^n)$ is outside S_1, so that $(\hat{\sigma}_{i-j_{01}}^n, \hat{r}_{i-j_{01}}^n) \in U(A,\varepsilon)\cap k(\Omega)$ will be defined as a data segment point.

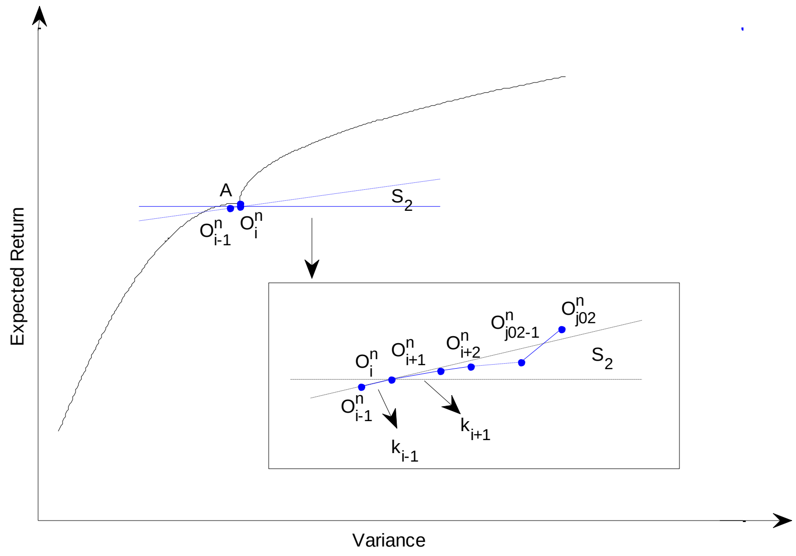


Figure 11. Convergence explanation of Type 2 segment point.

Let $A=(\tilde{\sigma},\tilde{r})$ be a real segment point of Type 2 on the CCMV efficient frontier. Let $k(x)=(\sigma(x),r(x))$ be the risk measure and the expected return measure with the weight vector x. According to Theorem 1, for any small neighborhood of A, $\int_{k^{-1}(U(A,\varepsilon))\cap\Omega} p(x)>0$. There exists a sample point $(\hat{\sigma}_i^n,\hat{r}_i^n)\in U(A,\varepsilon)\cap k(\Omega)$ (for n large enough, this is true).

Let: $S_2=\{(\sigma,r)|\sigma-\hat{\sigma}_i^n\le k_{i-1}\cdot(r-\hat{r}_i^n),\sigma\ge\hat{\sigma}_i^n,r\ge\hat{r}_i^n\}$, if $k_{i-1}\ge k_{i+j_2}$, which means $(\hat{\sigma}_{i+j_2}^n,\hat{r}_{i+j_2}^n)\in S_2,(j_2=1,2,\ldots)$ (see Figure 11).

Since A is a real segment point, there is a neighborhood in $(S_2)^c\cap k(\Omega_{i+1})$ so that $\int_{k^{-1}((S_2)^c\cap k(\Omega_{i+1}))} p(x)dx>\gamma>0$, where $\Omega_{i+1}=\{x|\tilde{\sigma}_i\le x'Gx\le\tilde{\sigma}_{i+1},\tilde{r}_i\le r'x\le\tilde{r}_{i+1},I'x=1,\sum_{i=1}^{n}|sign(x_i)|=K\}$. Thus let T_{j_2} represents the event that the sample point $(\hat{\sigma}_{i+j_2}^n,\hat{r}_{i+j_2}^n)$ $(j_2=1,2,\ldots)$ is not in the triangular area (S_2). Therefore, the probability of T_{j_2} has the property as follows:

$$\Pr(T_{j_2})\ge\delta>0 \quad (7)$$

Let T' represents the event that not a single sample $(\hat{\sigma}_{i+j_2}^n,\hat{r}_{i+j_2}^n)$ $(j_2=1,2,\ldots)$ of all n

portfolio samples is not in the triangular area (S_2). Then, the probability of T' satisfies the following formula: $\Pr(T') \leq (1-\delta)^{j_{02}} \to 0$. Consequently, in probability, there always exists a bounded number j_{02} such that $(\hat{\sigma}^n_{i+j_{02}}, \hat{r}^n_{i+j_{02}})$ is outside S_2 , so that $(\hat{\sigma}^n_{i+j_{02}}, \hat{r}^n_{i+j_{02}}) \in U(A, \varepsilon) \cap k(\Omega)$ is a data segment point.