APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind. **15**, 485–493 (1999)

NEW METHODS FOR ORDERING MULTIVARIATE DATA: AN APPLICATION TO THE PERFORMANCE OF INVESTMENT FUNDS

SERGIO ZANI*,†, MARCO RIANI AND ALDO CORBELLINI

Istituto di Statistica, Università di Parma, Parma, Italy

SUMMARY

This paper deals with the performance evaluation of investment funds. The goal is the monitoring of the profitability of the funds using a set of variables referred to as short- and medium-term performance, volatility and percentage of treasury bonds on total assets. The problem is the ordering of multivariate data and the search for the units lying far from the centroid. This question is related to the detection of multivariate outliers (Atkinson AC. Fast Very Robust Methods for the Detection of Multiple outliers. JASA 1994; **89**:1329–1339.; Riani M, Zani S. An Iterative Method for the Detection of Multiple outliers Metron 1997; **55**:101–117.

In this analysis we use a variation of the robust-bivariate boxplot for each pair of variables suggested in a previous paper first on the whole of the observations and then on each distinct category of funds (i.e. stock, mixed (balanced) and bond funds). The purpose is to split the units into a few subsets: those inside the inner region close to the centroid and those lying outside the outer contour (that is the most extreme observations). In the middle we can find groups of observations ordered according to their distance from a robust centroid. We apply our method first to the original variables and then to principal components and canonical variates. An analysis of longitudinal data is also considered. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: bivariate boxplot; *B*-spline; convex hull; multivariate ordering

1. INTRODUCTION

Usually, the performance of investment funds is analysed through linear econometric models. Recently, however, it has been recognized that linear models may be inadequate (e.g. Reference [1]). At present, the widespread belief is that the market efficiency hypothesis must be buried [2]. Attention has therefore been paid to non-normality or non-linearity and a large number of techniques have been suggested to cope with long-tailed distributions. However, models which use conditional heteroskedasticity (e.g. Reference [3]) have a good explanatory performance in-sample but a disappointing one for forecasting. Moreover, on the part of the operating units of financial companies, there is a lack of interest in statistical properties and asymptotic results.

* Correspondence to: S. Zani, Istituto di Statistica, Università di Parma, Parma, Italy.
† E-mail: zani@unipr.it

CCC 1524–1904/99/040485–09$17.50

Received June 1997
Revised May 1999

In this paper we propose a simple and appealing approach to the evaluation of the performance of investment funds which is based on the construction of the robust bivariate boxplot for each pair of variables. The suggested approach is easy to handle and can be conveniently used by operating units.

The structure of the paper is as follows: in Section 2 we initially recall briefly the steps of the construction of the bivariate boxplot [4]. This method shows some advantages relative to former proposals [5]. Then we apply this methodology to data on 105 Italian investment funds[‡]. First, we consider the whole of the observations and then each distinct category of funds. In Section 3 we show how the construction of a bivariate boxplot for each pair of variables can be a very useful tool for the visualization and the ordering of multivariate observations. In Section 4 we apply our method to the scores of principal components and canonical variates. Section 5 reports a longitudinal analysis of the performance of investment funds using bivariate boxplots. Section 6 concludes.

2. ROBUST BIVARIATE BOXPLOTS AND THE ORDERING OF BIVARIATE OBSERVATIONS

The construction of robust bivariate boxplot is based on the following steps.

Step 1: definition of the inner region

A natural and completely non-parametric way of finding a central region in $\mathbb{R}^2$ is through the use of the so-called convex hull peeling. Barnett [6] suggested that the most extreme group of observations in a multivariate sample are those lying on the convex hull (with those on the convex hull of the remaining sample, the second most extreme group, etc.). The output of the peeling is a series of nested convex polygons (hulls). We call the $(1-\alpha)$ per cent hull the biggest hull containing not more than $(1-\alpha)$ per cent of the data. The 50 per cent hull is the bivariate equivalent of the interquartile range. In order to ensure a smooth inner region we suggest drawing a *B*-spline curve inside the '50 per cent-hull' previously formed. This leads us to define the inner region as follows.

Definition

We call the 'inner region' that portion of the space which lies inside or on the boundary of the *B*-spline curve (hinge) superimposed on the 50 per cent hull.

This inner region is free from outliers and robust, but at the same time it keeps the correlation in the data and allows for differing spreads in the various directions.

Step 2: definition of the robust centroid

In this paper we have found the robust centroid using the arithmetic means of the observations inside the inner region. In this way, we exploit both the efficiency properties of the arithmetic

[‡] The data have been extracted from the Italian financial newspaper: II Sole 24 ore. Data available on request.

mean and the natural trimming offered by the hulls. Other criteria for the choice of the centroid are discussed in Reference [7].

Step 3: definition of the outer region

Once we have found a robust bivariate centre and a curve containing not more than half of the data (hinge), we must devise a criterion in order to build an outer contour which leaves outside a fixed proportion of the observations.

We suggest building the outer contour as a multiple of the distance of the hinge to the robust centre. In the detection of bivariate outliers, we find a multiple such that 1 per cent of the observations should lie outside the contour. Observations which lie outside this contour are referred to as atypical. As shown by Zani *et al.* [4], in order to obtain an outer contour which under the assumption of bivariate normality leaves outside a percentage of observations close to 1 per cent, we must multiply the distance of the hinge to the robust centre by a coefficient $l = 1.68$. This coefficient is independent of the parameter ρ of the bivariate normal distribution. In one dimension this would correspond to the independent use of the distances between the upper or lower quartile and the median in order to compute the upper or lower truncation points.

In this application, since the purpose is the detection of funds with good and bad multivariate performances, we can consider a variety of thresholds. Using a proportion of 10 per cent of external units we must multiply the distance of the hinge from the robust centre by $l = 0.86$. Using this coefficient we obtained the 90 per cent outermost contour reported in Figure 1. This threshold, however, can conveniently be decreased or increased in relation to the percentage of extreme observations one wants to detect. For example, in Figure 1 we have also considered a threshold of 75 per cent (which implies a coefficient $l = 0.43$).

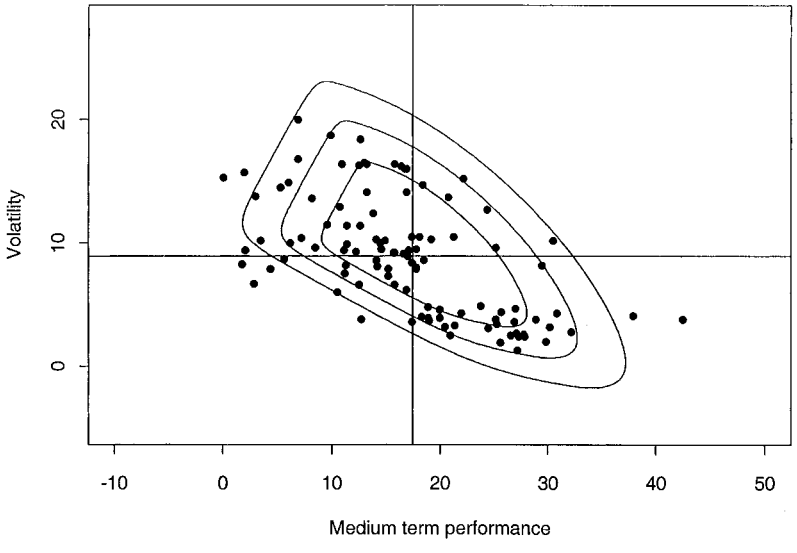

Figure 1. Robust centroid, inner region, 75 and 90 per cent outer contours of medium-term performance and volatility of 105 Italian investment funds (December 1996).

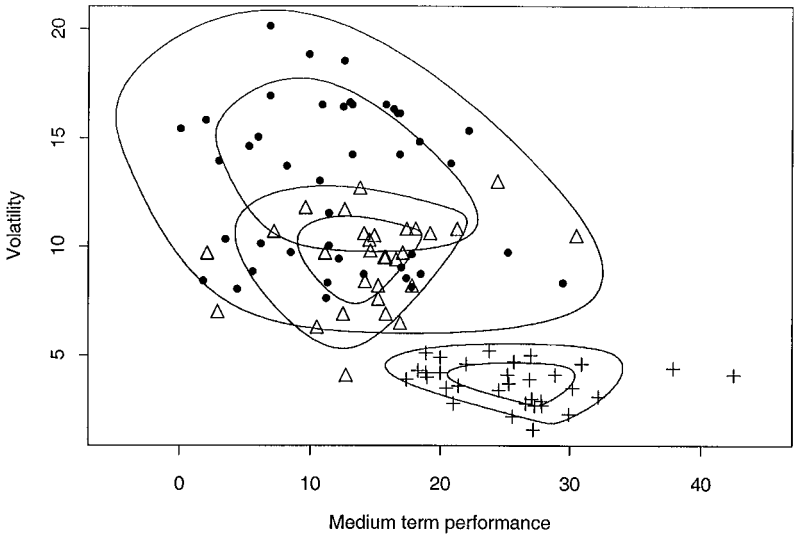

Figure 2. Inner region and 90 per cent outer contours of medium-term performance and volatility for each category of funds: ' + ' = bond funds, △ = mixed funds, ● = stock funds.

The above approach is an alternative to the one which aims to define quantiles for multivariate data [8]. Our method uses convex hull peeling and therefore is more closely related to that of Ruts and Rousseeuw (1996) [9]. These authors, using the notion of half-space depth, have proposed an exact algorithm for computing depth contours of a bivariate data set.

Figure 1 clearly shows that the shape of the different regions (which are non-parametric and smooth) adapts to the different spread of the data in different directions. The extreme funds are those outside the largest region. Of course, the best funds are those which combine the highest performance and the lowest volatility. From this plot we can also separate the funds which lie inside the hinge (i.e. close to the centre in terms of medium-term performance and volatility), from those which lie in the strip between the inner and the outer regions. This subgroup of units can be successively split into two additional subsets using an intermediate threshold of 75 per cent. In larger samples further thresholds could be used in order to obtain a more subtle ordering.

In Figure 2 we consider separately the three populations of funds. This plot shows that all bond funds have low volatility and high performance. Out of the outer region we can find two funds which have a medium-term performance much higher than the others. It is interesting to notice that there is no overlapping between the 90 per cent outer region of these funds and those of the other two populations. Figure 2 also shows that stock funds present high dispersion in terms of volatility and medium-term performance. Finally, the bivariate boxplot of mixed funds shows a high degree of overlapping with stock funds.

3. MULTIVARIATE VISUALIZATION

The two variables considered so far belong to a four-dimentional data set in which variables 3 and 4 are 'short-term performance' (12 months) and percentage of treasury bonds in total assets.

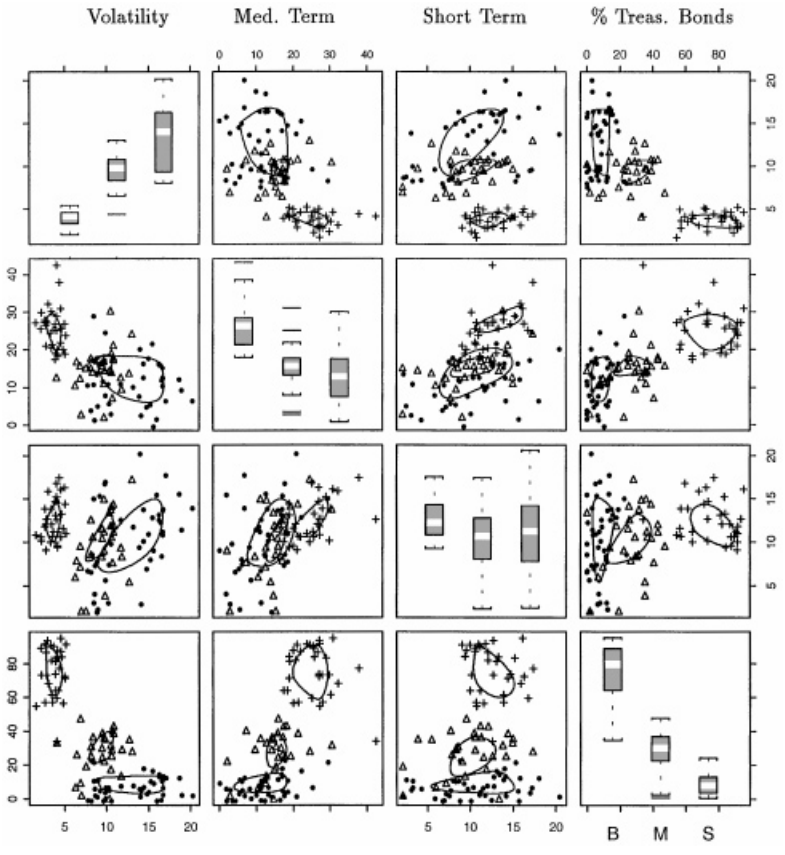

Figure 3. Scatter plot matrix of investment funds. Inner regions (50 per cent) for each category of funds superimposed on each diagram. Univariate boxplots for bond (B), mixed (M) and stock funds (S) added to the main diagonal.

A simple method of visualization of multivariate data is through the construction of bivariate boxplots for each pair of variables. Figure 3 reports the scatter plot matrix with superimposed 50 per cent inner regions for each population of funds. On the main diagonal we have also included the univariate boxplots for each category of funds. As this figure shows, while the units belonging to the cluster of bond funds tend to lie in a separate portion of the space, the 50 per cent inner regions of stock and mixed funds usually intersect showing a high degree of overlapping. This is also revealed by the univariate boxplots. The shape of the boxplots for the bond funds is always appreciably different from that of the other two categories of funds.

In conclusion, from the bivariate boxplots of each separate population we can extract useful information about: (a) location and spread of the different populations in the various directions; (b) the amount of overlapping; (c) the different behaviour of the subsets, (d) the atypical units, (e) the variables which best highlight the differences among the groups. This method provides useful information about the choice of the best investment funds; furthermore, it can be conveniently used for variable selection in discriminant analysis.

Table I. Proportion of variance and correlation coefficients for the first two principal components.

	First comp.	Second comp.
Proportion of variance	0.616	0.272
Volatility	− 0.814	0.494
Medium-term performance	0.891	0.212
Short-term performance	0.436	0.874
Percentage of treasury bonds	0.904	− 0.185

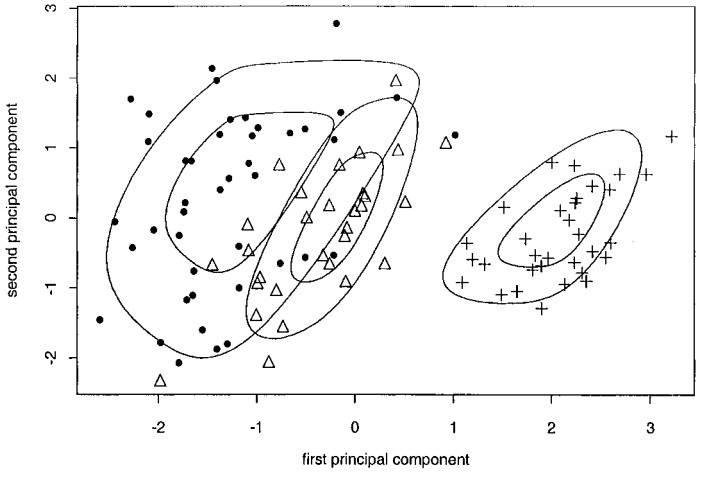

Figure 4. Data plotted against first two principal components. Bivariate boxplots with 90 per cent outer contours for each population have been superimposed on each subset of funds.

4. MULTIVARIATE ORDERING BASED ON PRINCIPAL COMPONENTS

The construction of bivariate boxplots can conveniently be applied not only to the original variables but also to the scores of derived variables such as principal components (PC) and canonical variates (CV).

Table I reports the results of PC analysis. The first two components explain 88.8 per cent of the variance of the original variables. The correlation coefficients between the variables and the PC, show that the first PC is positively correlated with medium-term performance and percentage of treasury bonds but negatively related to volatility. This latent variable reflects the evolution of the Italian financial market. In Italy during the years 1994–1996 there have been high interest rates on treasury bonds and an unsatisfactory evolution of the stock exchange. Therefore the bond funds, with a high percentage of treasury bonds and low volatility, showed the best performance. On the contrary, mixed and stock funds were characterized by high fluctuations and lower performances.

The second PC is positively correlated with short-term performance (12 months).

Figure 4, which reports the scores of the first two principal components, shows that bond funds have positive values for the first principal component and form a distinct cluster. The separation

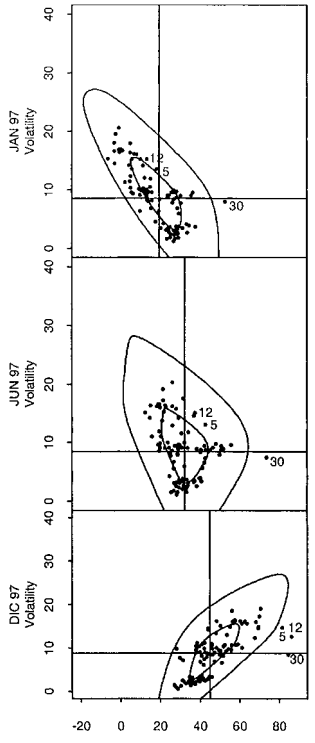

Figure 5. Bivariate boxplots of volatility and medium-term performance with 50 and 99 per cent contours for the whole set of funds. Top, central and bottom panels refer, respectively, to January 1997, June 1997 and December 1997.

between stock and mixed funds, however, is not very clear. The second component shows that the highest dispersion of short-term performance is for stock funds. The small loss of information provided by the projection of the four-dimensional units in the space of the first two PC implies that the units which are outside the outer regions constructed in this space may be considered as atypical for each category. This provides a natural ordering for multivariate data and must be considered as an alternative approach to those based on robust distances (e.g. Reference [10]).

We have also applied the technique of linear discriminant analysis. The results (not reported here) show that the first canonical variate (CV) explains 99.7 per cent of the differences among the groups. The correlation between the first PC and the first CV is 0.95. This implies that the direction in the four-dimensional space in which the projected points show maximum variance is essentially the same as that which best shows up differences among the 3 categories.

5. EXTENSIONS TO LONGITUDINAL DATA

Bivariate boxplots can also be used for the analysis of longitudinal data. In a three-way data matrix 'units × variables × times' it is possible to construct the inner region and the outer contour

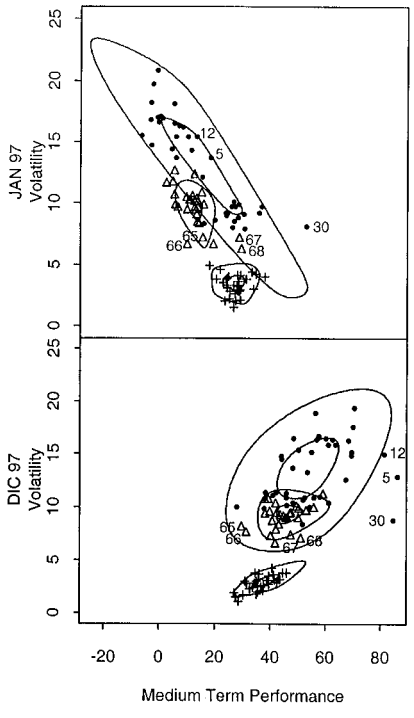

Figure 6. Bivariate boxplots of volatility and medium-term performance with 50 and 99 per cent contours for each category of funds. Top panel refers to January 1997, bottom panel to December 1997.

for each pair of (derived) variables in every time. Figure 5 shows the bivariate boxplots referred to a subset of 100 funds[§] for volatility and medium-term performance in the 3 periods January 1997, June 1997 and December 1997. The scatter diagrams show that the correlation between these two variables seems to change considerably during the 3 periods. This modification is clearly reflected by the change of shape of the outer contour. Furthermore, the superimposition of a bivariate boxplot enables us to state that unit 30 is an outperformer fund. Finally, it is important to notice that 2 other funds (unit 5 and unit 12) lie outside the outer threshold only in the last period. It is also interesting to analyse the evolution of the funds considering separately the different populations. For legibility on the printed page, in Figure 6 we only report the bivariate boxplots for the first and the last periods. This plot shows that the dispersion of stock funds is much bigger than that of the other two populations. The change of shape of the outer contour of bond funds denotes that, while in January 1997 the most profitable bond funds were those with the smallest volatility, in December 1997 high volatility generally seems to be associated with high performance. This reflects the general marked increase of the Italian and European stock exchange indexes in 1997. As regards mixed funds it is interesting to note that while in January 1997 their

[§] Unfortunately, we could not monitor all the 105 funds considered in the previous sections because 5 of them changed their category.

corresponding outer contour was only partially overlapping with that of stock funds, in December 1997 all mixed funds lie inside the outer contour which encloses stock funds. Finally, Figure 6 shows that bond funds, as expected, are very homogeneous in terms of volatility and medium-term performance.

In Figure 6, a few funds are identified by their number and this lets us monitor their dynamic movements. The superimposition of bivariate boxplots enables us to analyse what are the funds which remain inside the inner region and those which go out of it. This provides additional information about the evaluation of the global relative performance of the different funds.

6. CONCLUSIONS

The construction of bivariate boxplots for each pair of variables for data coming from different populations, can be a very useful instrument for extracting information about spread, location, and separation among the groups. These tools not only enable us to highlight potential outliers, but can help to select the variables in discriminant analysis and can provide useful information about the choice of multivariate transformations [11]. Furthermore, bivariate boxplots can also be drawn on derived variables such as principal components and canonical variates, provided that the projection gives a good approximation to the original space. The construction of bivariate boxplots on the first two principal components (canonical variates) provides a natural criterion for the ordering of multivariate data.

In this paper we have also considered the longitudinal analysis of the performance of the funds through bivariate boxplots. This technique can become a powerful instrument to judge the 'goodness' of the different funds, because it enables to analyze, in a completely non-parametric way, how far is the performance and volatility of each fund from the global one or from that of the corresponding category.

Finally, it is worthwhile to remember that the applicability of the suggested approach is not limited to the analysis of the stock exchange market, but can be naturally extended to a variety of practical situations.

REFERENCES

1. Kirman AP. Ants, rationality and recruitment. *Quarterly Journal of Economics*, 1993; 137–156.
2. Chiandotto B, Gallo GM. eds. *Quest of the Philosopher's Stone*, *Nonlinearity and Volatility in Financial Markets*, *Proceedings from the Satellite Meeting SIS*, *Exchange Rates and Financial Markets*: *Theoretical Models and Empirical Applications*, Imperia, 1994.
3. Bollerslev T, Engle RF, Nelson DB. ARCH models. *Handbook of Econometrics*. North-Holland: Amsterdam, 1994.
4. Zani S, Riani M, Corbellini A. Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis* 1998; **28**:257–270.
5. Goldberg KM, Iglewicz B. Bivariate extensions of the boxplot. *Technometrics* 1992; **34**:307–320.
6. Barnett V. The ordering of multivariate data. (with discussion). *JRSS*, Series A 1976; **139**: 318–339.
7. Riani M, Zani S, and Corbellini A. Robust bivariate boxplots and visualization of multivariate data. In: *Classification, Data Analysis and Data Highways*, Balderjahn I, Mathar R, Schader M. (eds), Springer: Berlin, 1998; 93–100.
8. Chaudhuri P. On a geometric notion of quantiles for multivariate data. *JASA* 1996; **91**: 862–872.
9. Ruts I, Rousseeuw PJ. Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis*. 1996; **23**:153–168.
10. Riani M, Zani S. Generalized distance measures for asymmetric multivariate distributions. In *Advances in Data Science and Classification*, Rizzi A, Vichi M, Bock H-H. (eds), Springer: Berlin, 1998; 503–508.
11. Atkinson AC, Riani M. Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis: the 1997 Hunter Lecture. *Environmetrics* 1997; **8**:583–602.

